Case Studies

Instructions

- This work should STRICTLY be done as a group work.
- Carefully read the case study.
- Discuss the answers to the questions in your groups.

Study 1: Data Pre-processing

Subtitle

Goal:

- To extract data from an online source.
- To pre-process data for cluster analysis and/or a classification model.

Approach:

- Using web scrapping, extract a noisy data set from an online source.
- Choose the appropriate techniques to deal with missing values and outliers - Choose the correct feature selection technique for cluster analysis.
- Choose the correct feature selection technique for a classification problem. - Perform feature normalization on the data set.
- Save the pre-processed data set in a CSV file.

Study 2: Pattern Mining

Supermarket shelf management

Goal:

- To identify items that are bought together by sufficiently many

customers. Approach:

- Search and retrieve a point-of-sale data from any source of your choice.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types etc.) - Process the point-of-sale data collected with barcode scanners to find any frequent dependencies among items.
- What conclusions can you make from the patterns extracted?

Study 3: Pattern Mining

Inventory Management:

Goal:

- A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.

Approach:

- Search and retrieve relevant data from any source of your choice.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types etc.) - Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Study 4: Classification

Direct Marketing

Goal:

- Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone

product. Approach:

- Search and retrieve relevant data from any source of your choice.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types etc.) - Choose the correct feature selection technique for a classification problem. - Perform feature normalization on the data set.
- Identify a feature in the data set that indicates if a customer "buys" or "doesn't buy" a product. If no feature exists, add it randomly for every customer i.e., {buy, don't buy}. The decision forms the class attribute.
- Collect various demographic, lifestyle, and company interaction related information about all such customers (i.e., type of business, where they stay, how much they earn, etc.) - Use this information as predictor features to build a classifier model.

Study 5: Classification

Fraud Detection

Goal:

- Predict fraudulent cases in credit card transactions.

Approach:

- Search and retrieve relevant data from any source of your choice.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types etc.) - Choose the correct feature selection technique for a classification problem. - Perform feature normalization on the data set.
- Use credit card transactions and the information on its account-holder as attributes (i.e., When does a customer buy, what does he buy, how often he pays on time, etc.)
- Label past transactions as fraud or fair transactions. This forms the class attribute. - Build a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

Study 6: Classification

Customer Attrition/Churn

Goal:

- To predict whether a customer is likely to be lost to a competitor.

Approach:

- Search and retrieve relevant data from any source of your choice.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types etc.) - Choose the correct feature selection technique for a classification problem. - Perform feature normalization on the data set.
- Use detailed record of transactions with each of the past and present customers, to find attributes (i.e., how often the customer calls, where he calls, what time-of-the day he calls

most, his financial status, marital status, etc.)

- Label the customers as loyal or disloyal.
- Build a classification model for loyalty.

Study 7: Classification

Sky Survey Cataloging

Goal:

- To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory). I.e., 3000 images with 23,040 x 23,040 pixels per image.

Approach:

- Search and retrieve relevant satellite data from a credible source.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types etc.) - Choose the correct feature selection technique for a classification problem. - Perform feature normalization on the data set.
- Segment the image.
- Measure image attributes (features) - 40 of them per object.
- Model the class based on these features.
- Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find.

Study 8: Clustering

Market Segmentation

Goal:

- Subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

Approach:

- Search and retrieve relevant data from any source of your choice.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types

etc.) - Choose the correct feature selection technique for cluster analysis.

- Perform feature normalization on the data set (if necessary).
- Collect different attributes of customers based on their geographical and lifestyle related information.
- Find clusters of similar customers.
- Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Study 9: Clustering

Anomaly detection

Goal:

- Detect significant deviations from normal behavior in credit card transactions of a group of people.

Approach:

- Search and retrieve relevant data from any source of your choice.
- Select relevant techniques to clean the data (missing values, outliers, wrong data types etc.) - Choose the correct feature selection technique for cluster analysis.
- Perform feature normalization on the data set (if necessary).
- Build a profile of the “normal” behavior (i.e., profile can be patterns or summary statistics for the overall population)
- Use the “normal” profile to detect anomalies (anomalies are observations whose characteristics differ significantly from the normal profile)

Study 10: Clustering

Document Clustering:

Goal:

- To find groups of documents that are similar to each other based on the important terms appearing in them.

Approach:

- Search and retrieve a group of books/documents your choice. Make sure they have meta-data tags.
- To identify frequently occurring tags in different documents. Form a similarity measure

based on the frequencies of different terms. Use this information cluster the books.

NB: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Study 11: Data Modelling

Data Warehouse

Goal:

- To design and develop a data warehouse using Pentaho.

Approach:

- Download and install Pentaho software. Create a data warehouse.
- Search and retrieve data stored in various forms (database, CSV, Excel etc.) - Perform ETL on the data retrieved and store it your data warehouse.
- Perform the OLAP operations using your data warehouse.

Study 12: Information Retrieval

Subtitle

Goal:

- To build an IR search model.

Approach:

- Search and retrieve a group of documents your choice.
- Pre-process the documents (i.e., construct “bag of words”, remove stop words, lemmatization etc.)
- Design an IR model that will help compute the most relevant words.
- Design a search model that will allow a user to enter a search phrase and your model will respond with a list of documents sorted in the order of relevance to the search phrase.