AIN429 Data Mining Laboratory

Assignment 1: Data Preprocessing
Date Issued: 25.10.2023
Date Due: 06.11.2023

Aim of the Experiment

In this assignment, we will focus on data preprocessing, which is the preliminary step of data mining. Our aim here is to make the data ready for use by performing the necessary data preprocessing. The assignment should be implemented as a single Jupyter Notebook. Your notebook should be clearly documented, using comments and Markdown cells to explain the code and results. At the end of this exercise, you will become familiar with preprocessing methods using Python libraries.

Data preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.

In the real-world data are generally:

- ***incomplete***: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.
- ***Noisy***: containing errors or outliers.
- ***Inconsistent***: containing discrepancies in codes or names.

Data Preprocessing Methods

- ***Data Cleaning:*** Data is cleansed through processes such as filling in missing values, smoothing the noisy data, identify outliers or resolving the inconsistencies in the data.
- ***Data Integration:*** Data with different representations/databases are put together and conflicts within the data are resolved. e.g., using multiple databases, data cubes, or files.
- ***Data Transformation:*** Data is normalized, aggregated, and generalized to gain a new set of data values.
- ***Data Reduction:*** This step aims to present a reduced representation of the data in a data Warehouse. So, reducing the volume but producing the same or similar analytical results.

Data cleaning

1. Fill in missing values (attribute or class value):
    - Ignore the tuple: usually done when class label is missing.
    - Use the attribute mean (or majority nominal value) to fill in the missing value.
    - Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
    - Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.
2. Identify outliers and smooth out noisy data:
    - Binning
        - Sort the attribute values and partition them into bins
        - Then smooth by bin means, bin median, or bin boundaries.
    - Clustering: group values in clusters and then detect and remove outliers (automatic or manual)
    - Regression: smooth by fitting the data into regression functions.
3. Correct inconsistent data: use domain knowledge or expert decision.

Data transformation

1. Normalization:
    - Scaling attribute values to fall within a specified range.
        - Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
    - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev
2. Aggregation: moving up in the concept hierarchy on numeric attributes.
3. Generalization: moving up in the concept hierarchy on nominal attributes.
4. Attribute construction: replacing or adding new attributes inferred by existing attributes.

Data reduction

1. Reducing the number of attributes
    - Data cube aggregation: applying roll-up, slice, or dice operations.
    - Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space
    - Principal component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data.
2. Reducing the number of attribute values

    - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
    - Clustering: grouping values in clusters.
    - Aggregation or generalization
3. Reducing the number of tuples
    - Sampling

Experiment

1. Download the dataset. The dataset will be shared on the Piazza group.
2. Perform any preprocessing steps that may be necessary to clean or filter the data.
3. Analyze, characterize, and summarize the dataset both before and after preprocessing using tables and graphs. Clearly explain and interpret analysis results.
4. Use a Scaling method and explain why you used it.
5. Examine the distribution of all featurettes and comment on the distribution in general.
6. Explain about the features on a correlation matrix or heatmap.
7. Summarize any insights which you gained from your analysis of the data.
8. You should submit your codes and report as a single Jupyter notebook.

Background information

We provide with you some basic tutorials for data preprocessing using Python.

- https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_data_preprocessing_analysis_visualization.htm
- https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/
- https://www.v7labs.com/blog/data-preprocessing-guide
- https://www.javatpoint.com/data-preprocessing-machine-learning

Notebooks for ML

- https://github.com/krasserm/machine-learning-notebooks
- https://www.kdnuggets.com/2016/04/top-10-ipython-nb-tutorials.html

Grading

You will present your projects during laboratory hours.

- Import dataset and summarize (%10)
- Preprocessing (%30)
- Visualization (%30)
- Report (%30)

REMARKS:

- Submission format:
    - studentID_name_surname_hw1.ipynb

- Your submission should be matched with the format above**. 10 point** penalty will be applied on mismatched submissions.
- You will use an online submission system to submit your experiments.
- https://submit.cs.hacettepe.edu.tr/ Deadline is 23:59. No other submission method (such as; CD or email) will be accepted.
- Do not submit any file via e-mail related to this assignment.
- The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms, or source codes.
- You can ask your questions through the course's Piazza group and you are supposed to be aware of everything discussed in the group.