Lab Manual

CSPC 303

Data Mining and Analytics

List of Experiments

Ex. 1 – Install Python programing or R programming IDE of your choice and explore it. Further download, install and explore the command line and GUI version of open source Weka software. Explore UCI Machine Learning repository (https://archive.ics.uci.edu/) and https://www.kaggle.com to be used as data source for your lab experiments.

Ex. 2 Download any two datasets of your choice comprising different types of attributes. Compute and analyze central tendency (mean, median and mode), dispersion (range, quartiles, interquartile range, variance and standard deviation) of different attributes and covariance and correlation matrix for the given datasets. Discuss your observations regarding which operation is logically apt for given attribute type and regarding characteristics of datasets which can be observed based on covariance and correlation matrix.

Ex. 3 Select two publically available datasets comprising different types of attributes viz. nominal, ordinal, interval-scaled and ratio-scaled. A dataset must comprise minimally 2 different types of attribute. Compute the proximity (similarity and/or dissimilarity) between data points using following metrics: Simple Matching Coefficient, Jaccard Coefficient, Cosine Similarity, Euclidean Distance, Manhattan Distance, Supremum Distance, and Correlation as similarity metric. Initially consider each attribute individually for populating corresponding proximity matrix then consider each data object as represented by a vector of mixed attribute types and compute the proximity matrix for your dataset. Discuss your observation regarding applicability of different metric and any pattern prevailing in your data.

Ex. 4 Select a dataset which have issues of missing values and noisy data points. This information can be checked from metadata or documentation provided with the dataset. Apply different missing values handing methods namely Ignore the tuple, Use a global constant to fill in the missing value, Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value, Use the attribute mean or median for all samples belonging to the same class as the given tuple, Use the most probable value to fill in the missing value on your datasets. Further, address the issue of noisy data points still pertaining in the datasets even after handling the missing values using Binning and Regression methods. Analyze the effect of different techniques on dataset in terms of statistical parameters such as central tendency and dispersion. (Later on this exercise will be extended in combination of any clustering and/or classification and/or association technique).

Ex. 5 Select a dataset which comprises numeric attributes of varying range. Apply different normalization techniques viz. Min-max normalization, z-score normalization, Decimal scaling on your datasets. Further, discretize the numeric attributes using Binning and Histogram analysis method. Analyze the effect of different techniques on dataset in terms of type of attributes, statistical parameters such as central tendency and dispersion and change in aptness of proximity metrics. (Later on this exercise will be extended in combination of any clustering and/or classification and/or association technique).