IS580 – Knowledge, Discovery and Mining

Assignment 2

Due Date: 8th of April

In this assignment, you are expected to apply dimensionality reduction techniques on a glass identification dataset ("assignment2_data1.csv"), and extract and evaluate association rules obtained from a banknote authentication dataset ("assignment2_data2.csv"). The details of the datasets are given in the next pages, however this information is not required for the analyses. The target variable of the first dataset is type of glass (*Type*), and of the second dataset is the class of the banknote (*Class)*.

Deliverables:

1. Report including stats, results of the analysis, graphs and comments as an Rmarkdown file named "yourID_assignment2.Rmd"
2. Data Files
    a. "yourID_ data1transformed.csv" and "yourID_data2transformed.csv" including transformed attributes
    b. "yourID_ Models.Rdata"

Required Packages for the Analysis:

ggplot2, beanplot, mRMRe, Fselector, discretization, arules, arulesViz

You may use different packages for visualization with stating the name of the package in your report.

Questions

A. Consider the data file "assignment2_data1.csv".
    1. Summarize general characteristics of the dataset with descriptive statistics, and interpret the results.
    2. Visualize the distribution of numeric attributes by using beanplots. Briefly comment on the graph.
    3. Apply principal component analysis (PCA) to the dataset and comment on the importance of the components.
    4. Visualize the relationship between numeric attributes and target variable *Type* by using scatter plots.
    5. Find the significance of attributes based on their correlation with the target attribute (*Type*). Select the most significant attributes. Interpret the results with using the information obtained by scatter plots (drawn in A.4).
    6. Find the significance of attributes on target variable by using Minimum Redundancy Feature Selection (mRMR). Select the most significant attributes. Interpret the results with using the information obtained by scatter plots (drawn in A.4).
    7. Discretize the numeric attributes with **mdlp** algorithm by supervising the process with *Type*. Save transformed data as "yourID_data1transformed.csv".

B. Consider the data file "assignment2_data2.csv".
    1. Summarize general characteristics of the dataset with descriptive statistics, and interpret the results.

2. Visualize the distribution of numeric attributes by using beanplots. Briefly comment on the graph.
3. Discretize the numeric attributes with **mdlp** algorithm by supervising the process with *Class*. Save transformed data as "yourID_data2transformed.csv".
4. Find the frequent itemsets, maximally frequent itemsets and closed frequent itemsets with **apriori** algorithm by setting minimum support as 0.1.
5. Extract the association rules from each itemset you have found in (B.2) which of right hand side only contains *Class* attribute and having confidence equal or greater than 0.8. State the number of rules you have found for each. Sort each rule sets by their confidence values. Explain the top two rules in text. Show their significance using support, confidence and lift measures.
6. Save all the models you developed in "yourID_ Models.Rdata" file.

Bonus question (10 pts): Use *hyperconfidence* (not taught during the class) as an interestingness measure. Interpret the results. How is hyperconfidence measure different from the other measures? You do not have to use all the other measures for comparison, just the ones you think best explain the associations among the data.

Assignment2_data1.csv Data Set Information:

Vina conducted a comparison test of her rule-based system, BEAGLE, the nearest-neighbor algorithm, and discriminant analysis. BEAGLE is a product available through VRS Consulting, Inc.; 4676 Admiralty Way, Suite 206; Marina Del Ray, CA 90292 (213) 827-7890 and FAX: -3189. In determining whether the glass was a type of "float" glass or not, the following results were obtained (# incorrect answers):

Type of Sample -- Beagle -- NN -- DA
Windows that were float processed (87) -- 10 -- 12 -- 21
Windows that were not: (76) -- 19 -- 16 -- 22

The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence...if it is correctly identified!

Attribute Information:

1. Id number: 1 to 214
2. RI: refractive index
3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
4. Mg: Magnesium
5. Al: Aluminum
6. Si: Silicon
7. K: Potassium
8. Ca: Calcium
9. Ba: Barium
10. Fe: Iron
11. Type of glass: (class attribute)
-- 1 building_windows_float_processed
-- 2 building_windows_non_float_processed
-- 3 vehicle_windows_float_processed
-- 4 vehicle_windows_non_float_processed (none in this database)
-- 5 containers
-- 6 tableware
-- 7 headlamps

Relevant Papers:

Ian W. Evett and Ernest J. Spiehler. Rule Induction in Forensic Science. Central Research Establishment. Home Office Forensic Science Service. Aldermaston, Reading, Berkshire RG7 4PN [Web Link]

Assignment2_data2.csv Data Set Information:

Source:

Owner of database: Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe, volker.lohweg '@' hs-owl.de)
Donor of database: Helene DÃ¶rksen (University of Applied Sciences, Ostwestfalen-Lippe, helene.doerksen '@' hs-owl.de)
Date received: August, 2012

Data Set Information:

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Attribute Information:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)