

WEKA ASSIGNMENT

I. Assessment Problem and Requirements For this coursework, the following five tasks are defined:

1) PCA Dimension Reduction (& Classification) Task Data file name: Dry_Bean_Dataset_Final.arff
Seven different types of dry beans were used, taking into account the features such as form, shape, type, and structure by the market situation. A total of 16 features; 12 dimensions, and 4 shape forms, were obtained from the grains. a) Plot the first two dimensions PCA b) Use the first n PCA which 95% variance covered as inputs to apply classification algorithms, present, TP Rate, FP Rate, Precision, Recall, F-Measure, Confusion matrix and Classification rate c) Compare the classification results from b) with the results with original features

2) Feature Selection & Decision Tree Task Data Set name: Diabetes_data_Final.arff There are 16 features, including gender and age demographic information, the target class is positive or negative. a) Use a feature selection algorithm to select the best feature set b) Using the selected features, run decision tree J48, illustrate the decision tree. Merge feature into a reasonable group in order to reduce the tree size, for example, merge age group as [0-24] as group 1, [25-34] as group 2, [35-44] as group 3, [45-54] as group 4, [55-64] as group 5, [65-74] as group 6, [75 above] as group 7 if it is necessary, compare the classification results above (with selected features) with the results with original features.

3) Classification Task: a) Apply the classification algorithms (at least two methods you have learned) to the datasets described in Tasks 1 and 2 (Dry_Bean_Dataset_Final.arff and Diabetes_data_Final.arff) and present the results b) Present why you choose these classification methods and give the conclusions

4) Cluster Task: Data Set name: tripadvisor_review_data_final.arff This is an unlabeled data set which can be used as cluster analysis, this data set is populated by crawling TripAdvisor.com. Reviews on destinations in 10 categories mentioned across East Asia are considered. Each traveller rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0), and an average rating is used against each category per user. Attribute 1 : Unique user-id Attribute 2 : Average user feedback on art galleries Attribute 3 : Average user feedback on dance clubs

Attribute 4 : Average user feedback on juice bars Attribute 5 : Average user feedback on restaurants
Attribute 6 : Average user feedback on museums Attribute 7 : Average user feedback on resorts
Attribute 8 : Average user feedback on parks/picnic spots Attribute 9 : Average user feedback on beaches
Attribute 10 : Average user feedback on theatres Attribute 11 : Average user feedback on religious institutions

a) Apply a clustering algorithm and plot the values of either “within cluster sum of squared errors” or “log likelihood” against the number of clusters, then determine an optimal number of the clusters (in general $k \leq 8$). b) Present cluster results c) Present the characteristics for each cluster by using descriptive statistics

5) Image Data Classification Task: You will collect your own image data file either from public resource or your own for the image classification (more than two classes) a) Describe the data set b) Use a classification algorithm to classify the image data set c) Validate the results

You are expected to complete the above tasks and submit a report of no more than 3,000 words explaining the completed tasks and the reference list does not count. The report should have a brief introduction where you summarise what you have achieved. It is not necessary to write a literature review. However, if you used a significant portion of code from other sources, you must reference it. You must also reference any scientific paper or other sources that you used for deciding which methods or parameters to use.

Note: It is suggested to use a 10-fold cross-validation method whenever applicable. It is advisable to use various types of drawings, screenshots, diagrams, summaries and tables in order to present the analysed problem in the most complete and transparent form for the report.