TDDD41/732A75: Clustering Lab

Goals

- Gain familiarity with the data mining toolkit, Weka
- Learn to apply clustering algorithms using Weka
- Understand outputs produced by clustering tools in Weka

Procedure

- **Dataset**

  In this lab we will work with a dataset from HARTIGAN (file.06). The file has been translated into ARFF, the default data file format in Weka. Download the dataset here. The dataset gives nutrient levels of 27 kinds of food. The mounts of energy, protein, fat, calcium and iron have been measured in a 3 ounce portion of the various foods.

  Press the **Preprocess**tab. Now Press the **Open** button and load *food.arff*. A description of each attribute can be seen by selecting the attribute from the list in the left hand side of the screen. The description appears in the right hand side of the screen. Press the **Edit** button, you can read and edit each instances.

  More info on Explorer-Preprocessing is available in the Explorer User Guide.

- **Cluster Data**

  Several clustering algorithms are implemented in Weka. In this lab we experiment with an implementation of K-means, **SimpleKmeans**, and an implementation of a density-based method, **MakeDensityBasedClusterer** in Weka.

  To cluster the data, click on the **Cluster** tab. Press the **Choose** button to select the clustering algorithm. Click on the line that has appeared to the right of the Choose button to edit the properties of the algorithm. You can find a detailed description of the algorithm by pressing the **More** button. Set the desired properties and press **OK**. In the Cluster mode, select "Use training set". Press the **Ignore** attributes button to specify which attributes should be used in the clustering. Click **Start**.

  Check the output on the right hand side of the screen. You can right click the result set in the "Result list" panel and view the results of clustering in a separate window. The result window shows the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters. Another way of understanding the characteristics of each cluster is through visualization. We can do this by right-clicking the result set on the left "Result list" panel and selecting "Visualize cluster assignments". You also can click the **Save** button in the visualization window and save the result as an arff file.

More info on Explorer-Clustering is available in the Explorer User Guide.

- SimpleKmeans

  Apply "SimpleKMeans" to your data. In Weka *euclidian distance* is implemented in SimpleKmeans. You can set the number of clusters and seed of a random algorithm for generating initial cluster centers. Experiment with the algorithm as follows:

  1. Choose a set of attributes for clustering and give a motivation. (**Hint**: always ignore attribute "name". Why does the name attribute need to be ignored?)
  2. Experiment with at least two different numbers of clusters, e.g. 2 and 5, but with the same seed value 10.
  3. Then try with a different seed value, i.e. different initial cluster centers. Compare the results with the previous results. Explain what the seed value controls.
  4. Do you think the clusters are "good" clusters? (Are all of its members "similar" to each other? Are members from different clusters dissimilar?)
  5. What does each cluster represent? Choose one of the results. Make up labels (words or phrases in English) which characterize each cluster.

- MakeDensityBasedClusters

  Now with MakeDensityBasedClusters, SimpleKMeans is turned into a denstiy-based clusterer. You can set the minimum standard deviation for normal density calculation. Experiment with the algorithm as the follows:

  1. Use the SimpleKMeans clusterer which gave the result you haven chosen in 5).
  2. Experiment with at least two different standard deviations. Compare the results. (**Hint**: Increasing the standard deviation to higher values will make the differences in different runs more obvious and thus it will be easier to conclude what the parameter does)

Submission

Submit a report describing your experiments. This should include the reasons why certain parameters/attributes/values were chosen, explanation of the procedure, and the explanation of the results. When explaining the experiments there is no need for explaining every step of the process (such as which buttons were clicked, etc). It is enough to say which algorithm was run and which arguements were used. In addition, do not copy-paste the full outputs from the tool. When presenting the results, present only those parts of the output which are relevant for what you are trying to explain. In cases where you are asked to compare different experiments, you should first compare the results and then try to reason about why the changes in the results occured (did not occur). Avoid writing statements without motivation, such as "Results are better/worse." In other words, explain in what respect they are better/worse and why this happened. Your lab report should also include the answers to **all** questions in the text.