**220CT - Data and Information Retrieval**

**Intelligent Systems for Big Data Lab 3. - Data Preprocessing**

**1. Data Preprocessing Questions**

a) What's noise? How can noise be reduced in a dataset?

b) Describe 3 different techniques to deal with missing values in a dataset. Explain when each of these techniques would be most appropriate.

c) What is data aggregation and why might it be used.

d) Why would you normalise and standardise a data set

f) Explain the data pre-processing techniques in detail?

Data Cleaning

Data integration

Data transformation

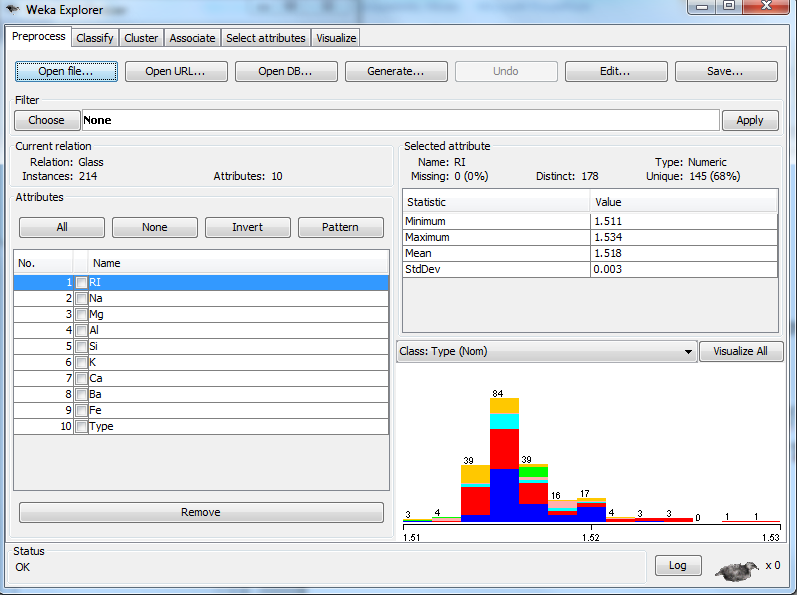
Data reduction

**2 Weka application for preprocessing**

If you have not attempted the second tutorial on data mining yet please try this first.

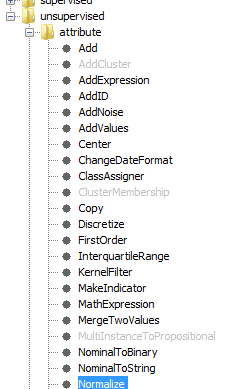
**Pre-processing**

Download the glass.arff data set from moodle to your computer. Open Weka, start the explorer window and choose the **preprocess tab**.



a) Using the **preprocessing tab** and opening the glass.arff file in word or wordpad. What information can you gain about the data set? How many different types of glass are there? How many attributes are used to determine which glasses is which? How many instances of each different type of glass are there? Which attitude has the largest mean? Which attribute has the largest standard deviation and what does this mean?

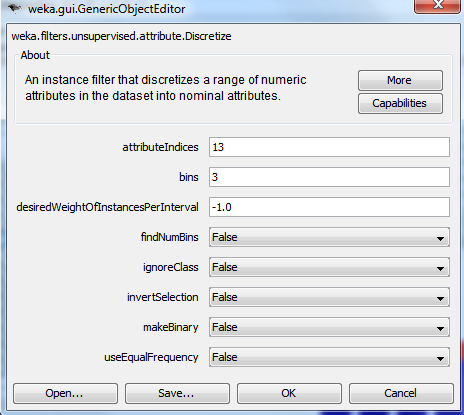
b) Select the **choose button** under **filters**, click on **unsupervised** and then **attribute** and **normalize**. Then click on **apply**.



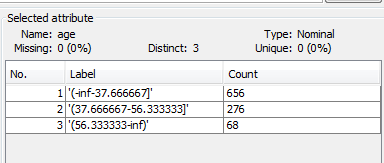
State what normalize does. Hint look at the max and min values for the attributes. Once you have done exploring this change, press the **undo button** to return the data back to the raw form.

c) Select **choose button** under **filters**, click on **unsupervised** and then **attribute** and **standardize**. Then click on **apply**. What does this filter do and what is impact on the statistical information provided for the attributes. Press **undo** when you have finished looking at this data.

d) Download the credit.arff data from moodle and open it in Weka. We can use WEKA to perform discretization on the "age" into distinct categories. In this example, we divide each of age into 3 categories. To do this in **filters** select the **choose button, unsuperized, attribute, Discretize.** Right click on **Discretize** and select **show properties** set **attributeIndices** to **13** and **bins** **3**. Attribute 13 is age and bins is in the number a categories. Click **OK and Apply**.

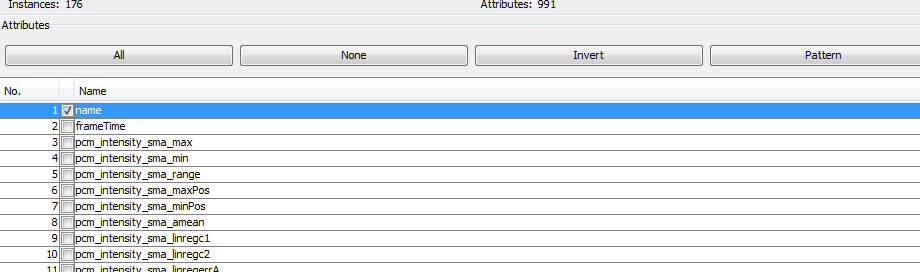


The three categories for age can be seen below.

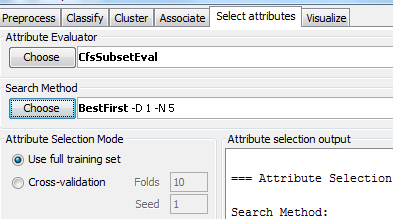


Now you try to following the same process to create 4 categories (bins) for credit\_amount.

d) Download the emotion.arff from moodle and open it in Weka. This dataset has over 900 features and has features extract from different utterances with different emotions. Weka offers approaches to reduce the number of features by determining which features best represent the classes (emotions). In this case these classes are different emotions. First click the box next to the **name attribute** and click **remove**.



Click the **Select Attribute tab** and then on the Search Method select **Best Fit** and for the Attribute Evaluator **CfsSubsetEval**. If you really want to know what these are right click on the words and select **show properties**.



Press **start**. How many attribute does Weka feel are the best to perform emotion recognition (See right panel). Give a few examples of these features. You might consider creating a MLP (see last times tutorial) using only the best attributes identified by WEKA (Clue use the select all and then unclick the attributes you wish to use and remove others).