Lab Exercise 6

Classification of Incomplete Data

Measurement data used in the classification process rarely occur in pure form, i.e. where the data are free of various types of noise and measurement errors. Measurement data is often damaged or missing individual values. When developing machine learning models, it is important to identify, mark and capture the missing data in order to get the best possible results from the algorithm.

The aim of this lab is to evaluate the performance of three data classification algorithms, namely:

- Naive Bayes,
- SVM,
- Artificial Neural Networks,

to classify incomplete measurement data and to **use WEKA package for filtration and cleansing of incomplete measurement data**.

The following instructions will answer your questions:

I. How to mark missing values in a dataset?
II. How to delete records with missing values from the data set?
III. How to assign missing values?

For this purpose, we will use a data set:

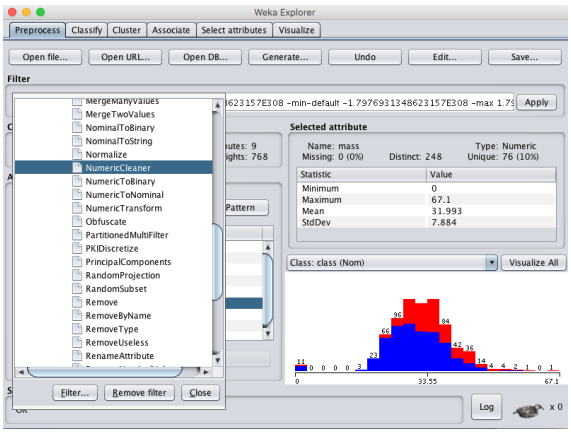
Pima Indians (https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes).

This is a classification problem where each record represents medical data for one patient and the task is to predict whether a patient will develop diabetes within the next five years. You can also access this set of data in the Weka installation, in the date directory in a file named diabetes.arff

I. How to mark missing values in a dataset?

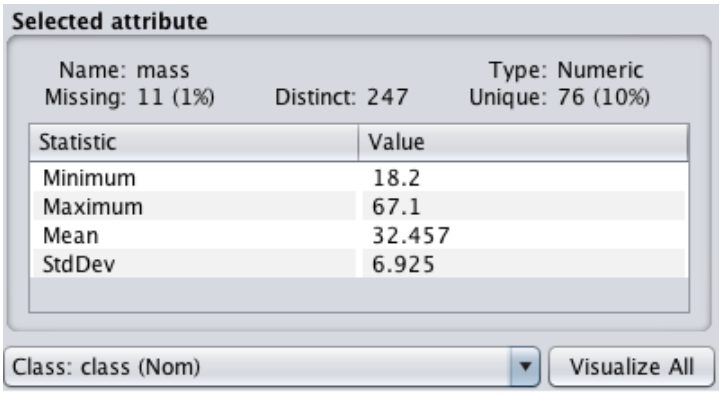
The Pima Indians dataset is a good example of searching for missing data. Some attributes such as blood pressure (pres) and Body Mass Index (mass) have zero values that are biologically impossible. These are examples of damaged or missing data that must be marked manually. The WEKA allows you to mark missing values using the *NumericalCleaner* filter. The following instructions show how to use this filter, mark the 11 missing values in the Body Mass Index attribute (mass).

1. Start Wekę and choose **Explorer**
2. Load *Pima Indians* data set.
3. Click “Choose” in tab Preproces -> Filter and choose *NumericalCleaner*, which you will find in *unsupervized.attribute.NumericalCleaner*.



4. Select filter in order to move to its configuration
5. Set *attributeIndicies* to 6 (index for attribute “mass”)
6. Set *minThreshold* to be 0.1E-8 (close to zero), which is the minimum possible value of this attribute.
7. Set *minDefault* to be NaN, which represents an unknown value and will replace all values below the threshold value, and click "OK" and then „*Apply*” in order to apply the filter to the dataset at hand.

Select the "mass" attribute in the attribute panel and familiarize yourself with the details of the selected attribute. Note that the 11 attribute values that have been set to 0 are now marked as missing.



In this example, we have marked values below the threshold value as missing.

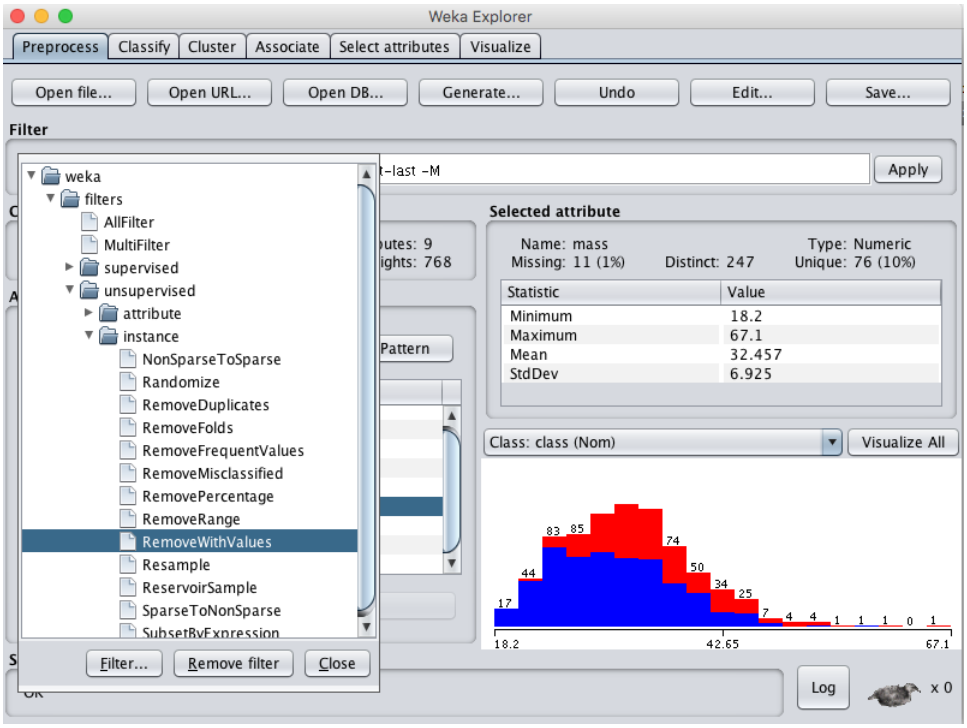
You can also easily mark them by assigning a different numeric value. You can also mark values in the numerical range as missing, i.e. when they are between the upper and lower threshold value.

Then, let's look at how we can delete records with missing values from our database.

II. II. How to delete records with missing values from the data set?

A simple way to deal with missing data is to delete those occurrences that have one or more missing values. In Weka this can be done using the *RemoveWithValues* filter. Following with the example above, you can delete the missing values as follows:

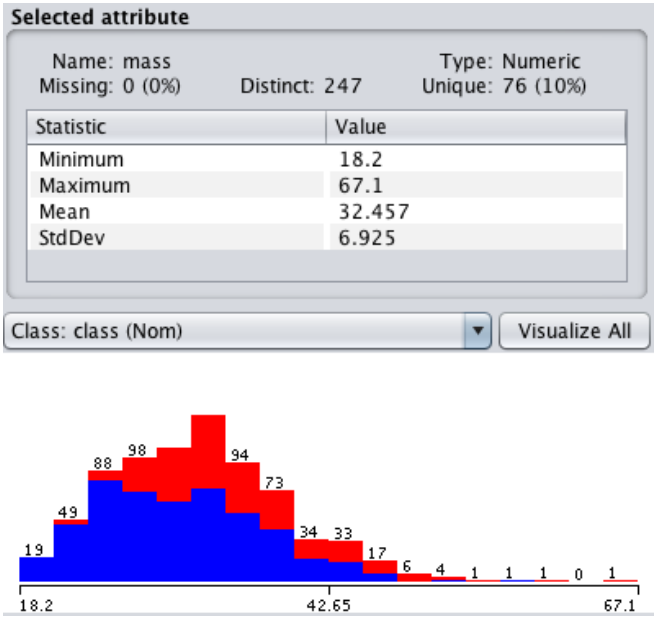
1. In the Filter selection window, select the *RemoveWithValues* filter and click „*Choose*" button.



2. Select a filter to configure it.
3. Set *attributeIndicies* to 6 (index for attribute "mass")
4. Set *matchMissingValues* to *True*
5. Press "OK", in order to use filter configuration.
6. Press "Apply", in order to use selected filter.

Select the "mass" attribute from the attributes section and view the details of the selected attribute.

Please note that the 11 attribute values that were marked as "Missing" in the previous step have now been removed from the data set.

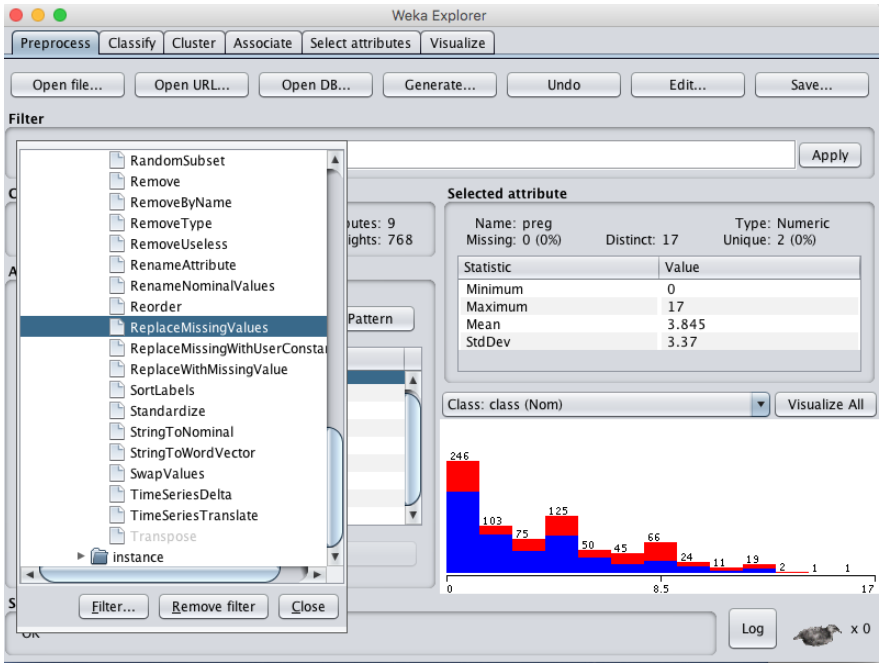


III. How to assign missing values?

However, records with missing values do not have to be deleted, and it is possible to replace the missing values with some other predetermined value. Such solution to the problem of missing values is called an assignment.

The frequently used method of assigning missing values uses the average values of the distribution of a given variable. This can be easily done in Weka using the *ReplaceMissingValues* filter. Following with the first example, you can assign the missing values in the following way:

1. In the filter selection window, select *ReplaceMissingValues* filter and click "*Choose*".



2. Click 'Apply' to apply the filter to the data set.

Click on the "mass" attribute in the attributes section and view the details of the selected attribute. Note that the 11 attribute values, which in the first example were marked as "Missing", were set to the average values of the distribution.

Tasks to complete (Report)

1. Download the Pima Indians collection (included in the WEKA package).
2. Prepare the data to be loaded into WEKA.
3. Launch WEKA and build 3 independent classification models to predict whether a patient will develop diabetes within the next five years. These models should be built on the basis of the following algorithms:
    a. Naive Bayes classifier– **NaiveBayes,**
    b. Support Vector Machine– **SMO,**
    c. Artificial Neural Network– **MultilayerPerceptron,**
4. For each of the above algorithms it is necessary to build 3 models (9 models in total):

    1) Using a data set with missing values
    2) Using a data set with deleted missing values
    3) Using a data set with substituted missing values

    In particular, the report should discuss the impact of missing/substituted values on the learning and prediction process and which algorithms have coped with the missing values and which have not.
5. By changing classifier parameters, create a satisfactory classification algorithm by analysing the Confusion Matrix and calculating and comparing parameters such as:
    a. Accuracy
    b. Sensitivity
    c. Specificity – What is the name of this parameter in WEKA?
    d. False-positive rate
6. By analyzing the above parameters and the ROC (Receiver Operating Characteristic) curve, compare the created algorithms and sort them in the best to the worst order.
7. Write the experiment report containing:
    a. Part describing principles of operation for each algorithm (use WEKA documentation and online materials)
    b. Describe data sets (how many / what are the attributes, what does the data collection describe, comment on data diversity and distribution as well as add any other comments/descriptions about the data set that you consider important).
    c. Present results from the classification process presenting each decision tree / decision table for each data set.
    d. Discuss the results as well as present the conclusion and the summary of the experiment

The report should be sent by email in **ONE** pdf file. When naming the file please use the following naming convention: **DM_LAB6_name_surname.pdf.** An email with the file should be sent to the email address of the lecturer and titled: **DM_LAB6_name_surname**