CS286 Lab 3: Data pre-processing and attribute selection in WEKA

K. A. Tarnowska

14-Sep-2020

Contents

1 Introduction **2**
1.1 Data pre-processing [5] 2
1.1.1 Data pre-processing in WEKA [4] 3
1.2 Attribute selection [3] 3
1.2.1 Attribute selection in WEKA [4] 4

2 Tasks **4**

3 Deliverables **7**

4 Grading **7**

5 Self-check **7**

1 Introduction

In this Lab, you will perform data pre-processing and feature selection in WEKA on the chosen dataset.

Pre-requisites This lab assumes completion of Lab 1 and Lab 2, that is, you should have a dataset identified and successfully loaded into WEKA.

1.1 Data pre-processing [5]

Data in the real-world is often *dirty*; that is, it is in need of being cleaned up before it can be used for the desired purpose. This is often called **data pre-processing**. Below are some factors that indicate that data is not clean or ready to process:

- **Incomplete** - when some of the attribute values are lacking.
- **Noisy** - when data contains errors or outliers.
- **Inconsistent** - when data contains discrepancies in codes or names.

The most important tasks involved in data pre-processing are:

- **Data Cleaning** - there are three key methods which describe how data may be "cleaned", better organized, or scrubbed of potentially incorrect, incomplete, or duplicated information:
  - **Data Munging** - converting the data to a format that is suitable for a computer to understand. While there is no specific scientific method, the approaches to take are all about manipulating or wrangling (or munging) the data to turn it into something that is more convenient or desirable.
  - **Handling Missing Data** - strategies to combat missing data include ignoring the record (the object), using a global constant to fill in all missing values, imputation, inference-based solutions (Bayesian formula or a decision tree), etc.
  - **Smooth Noisy data** - sometimes data is not missing, but it is corrupted for some reason, i.e., due to faulty data collection instruments, data entry problems, or technology limitations. While there is no single technique to remove noise, or smooth out the noisiness in the data, there are some steps to try. First, outliers should be identified and removed. Second, inconsistencies in the data should be resolved.
- **Data Integration** - data from various sources dan be integrated, that is, combined into a single file or a database from multiple sources.

- **Data Transformation** - data should be transformed so it is consistent and readable by a system. The following five processes can be used for data transformation:
  - **Smoothing** - remove noise from data.
  - **Aggregation** - summarization, data cube construction.
  - **Generalization** - concept hierarchy climbing.
  - **Normalization** - scaled to fall within a small, specified range and aggregation. Some of the techniques for normalization are:
    - Min-max normalization.
    - Z-score normalization.
    - Normalization by decimal scaling.
  - **Attribute or feature construction** - new attributes constructed from the given ones.
- **Data Reduction** - a process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained. The goal of **dimensionality reduction** is to identify which features to remove or collapse to a combined feature.
- **Data Discretization** - often the data collected from processes is continuous, i.e. temperature, company's stock price. The processing of converting (or mapping) continuous values into parts is called *discretization*. To achieve discretization, the range of continuous attributes are divided into intervals. For example, we could decide to split the range of temperature values into cold, moderate, and hot, or the price of company stock into above or below its market valuation.

More on data pre-processing techniques can be found in [5].

1.1.1 Data pre-processing in WEKA [4]

The data pre-processing algorithms in WEKA can be performed under *Preprocess* tab in WEKA *Explorer*. They are called *filters*. Clicking *Choose* (near the top left) in the *Preprocess* panel gives a list of filters. The filtering algorithms in WEKA are divided into *supervised* and *unsupervised*. Within each type, there is a further distinction between *attribute filters*, which work on the attributes in the datasets, and *instance filters*, which work on the instances. Refer to Chapter 8 in [3] and documentation available in WEKA to learn more about a particular filter.

1.2 Attribute selection [3]

Because of the negative effect of irrelevant attributes on most machine learning schemes, it is common to precede learning with an attribute selection (or feature selection) step that eliminates most but the most relevant attributes. Reducing

the dimensionality of the data by deleting unsuitable attributes improves the performance of learning algorithms. More importantly, dimensionality reduction yields a more compact, more easily interpretable representation of the target attribute (decision) and focusing the analyst's attention on the most relevant variables. Sometimes manual methods may be applied to eliminate (remove) irrelevant attributes based on domain knowledge. When this knowledge is not available, automatic methods can be helpful.

1.2.1 Attribute selection in WEKA [4]

The *Select attributes* panel gives access to several methods for attribute selection. These involve an attribute evaluator and a search method. Both are chosen and configured with the object editor. You must also decide which attribute use as the decision attribute (a "class"). Attribute selection can be performed using the full training set or using cross-validation (done separately for each fold, and displaying in how many folds-each attribute was selected). Alternatively, you can access attribute selection form the *Preprocess* panel, using *supervised*, *attribute*, *Attribute Selection* filter. A complete list of attribute selection methods implemented in WEKA and their descriptions can be found in Chapter 8 in [3] and in the documentation available within WEKA.

2 Tasks

Work with the dataset chosen for Lab 1 and Lab 2.

1. Find and describe any incomplete, noisy, or inconsistent data you found in your dataset.
2. Apply at least one pre-processing technique in WEKA (i.e. discretization) and provide screenshots of the results (see Figure 1 as an example). The method used for the pre-processing technique should be described (i.e. if you are applying discretization, describe the method you are using, i.e. unsupervised discretization, entropy-based discretization, or other).
3. Perform attribute selection, using and comparing at least three methods implemented in WEKA. Provide screenshots from WEKA of the result of each method (see Figures 2 and 3) as examples. Each method should be:
   - justified - why you chose this method;
   - described - what algorithm it implements and how it works;
   - parametrized - what parameters were chosen for the method;
   - screenshotted - provide a screenshot of the results;
   - commented on - comment and discuss the results.

Refer to [1] as a case study for data pre-processing and feature selection for tinnitus datasets and [2] as a case study for data pre-processing and attribute selection on the customer survey dataset.

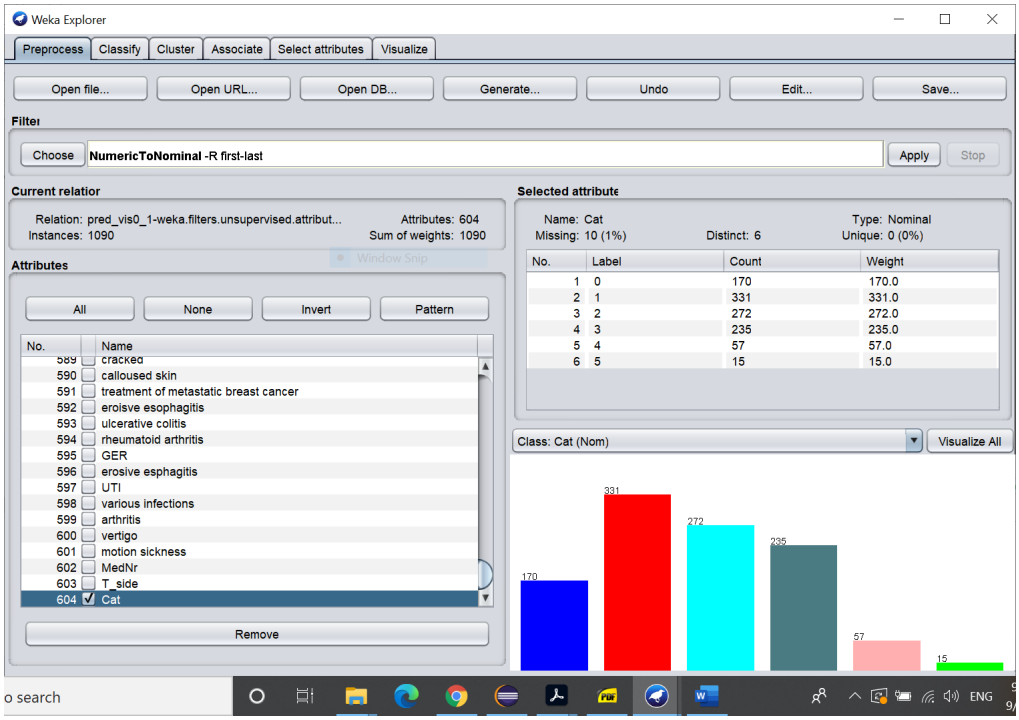

Figure 1: Example of data pre-processing on tinnitus datasets - converting numeric decision attribute (Category) to nominal using WEKA filter.

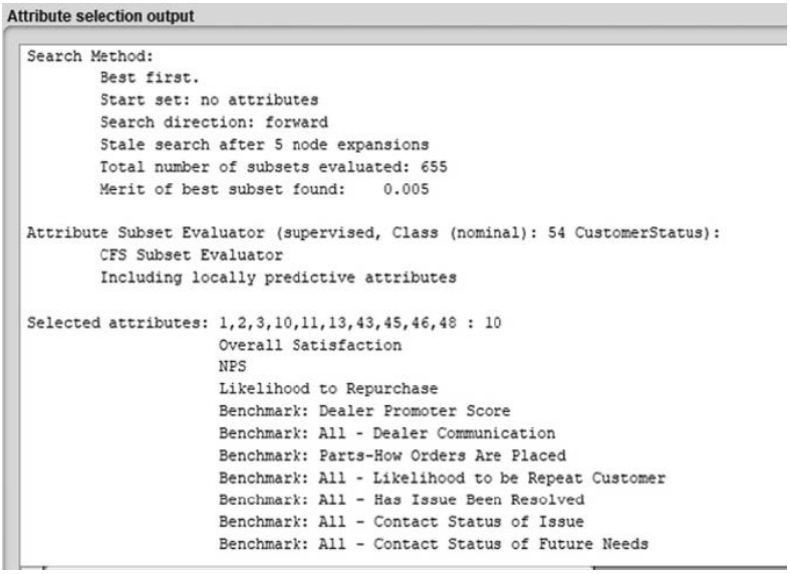

Figure 2: Example of attribute selection using the best first method in WEKA for the customer attrition problem.

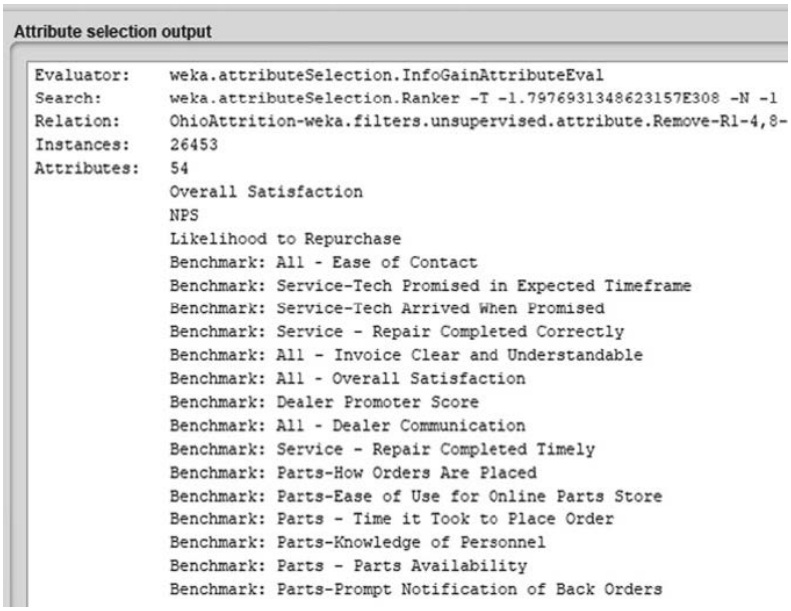

Figure 3: Example of attribute selection using information gain method in WEKA for the customer attrition problem.

3 Deliverables

Lab report uploaded on Canvas (pdf preferred) with embedded screenshots.

4 Grading

The lab will be graded according to the rubric in Table 1.

Table 1: Grading rubric

Rubric	Criteria	Points
1	Described noise, inconsistency, or incompleteness found in the dataset	3
2	Chose, described, and performed *at least one* data pre-processing technique (screenshot provided)	3
3	Chose, described, performed, compared, and commented on *at least three attribute selection* methods (screenshot of each provided)	4
Total		**10**

5 Self-check

Answer the following questions:

1. What are the common problems with *raw ("dirty") data*?
2. What are the most important tasks involved in *data pre-processing* and what do they mean?
3. What are the five processed that can be used for *data transformation*?
4. How is data pre-processing performed in WEKA?
5. Describe at least one pre-processing technique (*filter*) available in WEKA.
6. What is *attribute selection* and what is its purpose?
7. Describe at least three methods for attribute selection available in WEKA.

References

[1] Tarnowska KA, Ras ZW, Jastreboff PJ. Chapter 6.1: Initial Feature Development, Chapter 6.2.2: Feature Selection; Chapter 6.3.1-6.3.2: Pharmacology Data Analysis; Pivotal Features Development in: Decision Support System for Diagnosis and Treatment of Hearing Disorders. The Case of Tinnitus. Studies in Computational Intelligence. Springer, 2017

[2] Tarnowska KA, Ras ZW, Daniel L. Chapter 5.2: Data Preparation; 9.5.2: Attribute Selection in: Recommender System for Improving Customer Loyalty. Studies in Big Data. Springer, 2019

[3] Witten IH, Frank E and Hall MA. Data transformations, Chapter 8, in: Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Morgan Kaufmann. 2017

[4] Witten IH, Frank E and Hall MA. WEKA workbench, Appendix B, in: Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Morgan Kaufmann. 2017

[5] Data cleaning and pre-processing presentation: https://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf Accessed September 10, 2020