**Classification of Cancer Data using Weka**

Harris Sainulabdeen  
BU MET CS699 Spring 2017

# Part 1: Background and file descriptions

* **Background:**

The WHO Mortality Data base comprises deaths registered in national vital registration systems, with underlying cause of death as coded by the relevant national authority. Underlying cause of death is defined as “the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury” in accordance with the rules of the International Classification of Diseases.

The database contains number of deaths by country, year, sex, age group and cause of death as far back from 1950. Data are included only for countries reporting data properly coded according to the International Classification of Diseases (ICD).

The data available on this web site comprise deaths registered in national vital registration systems, with underlying cause of death as coded by the relevant national authority. These data are official national statistics in the sense that they have been transmitted to the World Health Organization by the competent authorities of the countries concerned. Each Member State reports population data along with their mortality data, for the population covered by the death registration system. Where this is a subset of the national population, the data is labelled accordingly in the WHO Mortality Database, e.g. Brazil (North and North-east) or Paraguay (reporting areas). However, the completeness of death registration may also be less than 100% for the specified registration population.

* **Dataset:**

The data set is distributed in compressed form. All the files are in ASCII format only (csv files with comma separated values)

* **Description of the data files**:

There are 4 files, named

* **MortIcd07.zip**: mortality data aggregated according to the Lists A and B of the 7th revision of the ICD. Total records: 281,749 (excluding first row of field names)
* **MortIcd08.zip**: mortality data aggregated according to the Lists A and B of the 8th revision of the ICD. Total records 380,627 (excluding first row of field names).
* **MortIcd09.zip**: mortality data aggregated according to the Basic Tabulation List (BTL) of the 9th revision of the ICD. Total records 894,099 (excluding first row of field names)
* **MortIcd10.zip**: mortality data according to the detailed 10th revision of the ICD, either with 3 characters or 4 characters ICD 10 codes. Total records 2,505,157 (excluding first row of field names).

# Part 2: Project Outline

* **Objective**

To present the analysis of publicly available WHO 57 years mortality data set with focus on two types of widely prevalent cancer; Lung Cancer and Leukemia.

* **Outcome**

The data was pre-processed using R and Weka, and classified using Weka. Four different classification algorithms were compared to each other in terms of performance. The models were found to classify Lung Cancer vs. Leukemia to varying degrees of accuracy. Overall, these two cancer types can be classified based on age of individuals at death.

* **Project outline**

The dataset can be useful for research into global trends and causes of death. In this project, the focus is on:

1. Reducing the number of attributes across the various data sets using attribute selection algorithms.
2. Preprocessing/preparation/transformation of the data sets so that a reduced number of tuples are used in the final analysis.
3. Classify Mortality data according to cause of cancer. Focus is only on 3 types of Cancer:
   1. 'A050' - Malignant neoplasm of trachea, bronchus and lung
   2. 'A058' - Leukaemia and aleukaemia

# Part 3: Data Pre-Processing

Here is the structure of Mortality Data file:

| **Column name** | **Content** |
| --- | --- |
| Country | Country code – see file “Country\_codes.zip” |
| Admin1 | Specified region/Category pertinent to each country– see Annex Table 2 below. If both fields 'Admin1' and 'Subdiv' are blank, data reported refer to the country. |
| Subdiv | Category of data – see Annex Table 2 below. If both fields 'Admin1' and 'Subdiv' are blank, data reported refer to the country. |
| Year | Year to which data refer |
| List | List of ICD revision used – see Annex Table 2 below. |
| Cause | Cause of death – For details consult Part 2 below or ICD publications |
| Sex | 1 male, 2 female and 9 sex unspecified |
| Frmat | Age-group format for breakdown of deaths at 0-95+ yrs – see Annex Table 1 below for details |
| IM\_Frmat | Age format for breakdown of infant deaths (0 year) – see Annex Table 1 below for details |
| Deaths1 | Deaths at all ages |
| Deaths2 | Deaths at age 0 year |
| Deaths3 | Deaths at age 1 year |
| Deaths4 | Deaths at age 2 years |
| Deaths5 | Deaths at age 3 years |
| Deaths6 | Deaths at age 4 years |
| Deaths7 | Deaths at age 5-9 years |
| Deaths8 | Deaths at age 10-14 years |
| Deaths9 | Deaths at age 15-19 years |
| Deaths10 | Deaths at age 20-24 years |
| Deaths11 | Deaths at age 25-29 years |
| Deaths12 | Deaths at age 30-34 years |
| Deaths13 | Deaths at age 35-39 years |
| Deaths14 | Deaths at age 40-44 years |
| Deaths15 | Deaths at age 45-49 years |
| Deaths16 | Deaths at age 50-54 years |
| Deaths17 | Deaths at age 55-59 years |
| Deaths18 | Deaths at age 60-64 years |
| Deaths19 | Deaths at age 65-69 years |
| Deaths20 | Deaths at age 70-74 years |
| Deaths21 | Deaths at age 75-79 years |
| Deaths22 | Deaths at age 80-84 years |
| Deaths23 | Deaths at age 85-89 years |
| Deaths24 | Deaths at age 90-94 years |
| Deaths25 | Deaths at age 95 years and above |
| Deaths26 | Deaths at age unspecified |
| IM\_deaths1 | Infant deaths at age 0 day |
| IM\_deaths2 | Infant deaths at age 1-6 days |
| IM\_deaths3 | Infant deaths at age 7-27 days |
| IM\_deaths4 | Infant deaths at age 28-364 days |

**Pre-Processing in R**

* All 4 data files were imported into R, converted into data frames.
* These data frames were then filtered to include only Leukemia and Lung Cancer as the causes of death.
* Columns were re-arranged to place the ‘Cause of Death’ column at the end for ease of classification in Weka.
* The cleaned data frames were then were written into CSV files.
* The CSV files were in turn loaded into Weka, and saved as .arff files.
* All subsequent pre-processing was done in Weka

**Pre-Processing in Weka**

|  |  |  |
| --- | --- | --- |
| MortIcd07.csv | Created MortIcd7.arff | Converted csv to arff in Weka |
| MortIcd08.csv | Created MortIcd8.arff | Converted csv to arff in Weka |
| MortIcd09.csv | Created MortIcd9.arff | Converted csv to arff in Weka |
| MortIcd10.csv | Created MortIcd10.arff | Converted csv to arff in Weka |

Out of the data CSV files, filter out only data related to various types of cancer. The result is cancer data stored in 4 files for years 1950-2012.

**Further Steps**

1. Normalize data, using the uniform Cancer Codes. There are more than 200 of them but want to reduce them to 3 common codes across the four files.
2. Use Attribute Selection Algorithms in Weka, against CancerData7. This is the training data
3. Apply the model on CancerData8. This is the test data

# Part 4: Classification

**Model Building:**

In this project the focus is only on Cancer related mortality in the entire data set. The data set is made up of a period data, including many years. There will not be filtering out of any years’ data. Similarly, all population age groups will be retained. Out of 255 different causes, only 3 will be selected. They are:

* + - 'A050' - Malignant neoplasm of trachea, bronchus and lung
    - 'A058' - Leukaemia and aleukaemia

**Training Data** – Selected 1950-1972 Dataset for training each classification algorithms owing to two reasons:

* Primarily this dataset had a manageable number of records (3281) .
* Selection based on chronological order, since the test data was newer (1968 to 2008) while the training data was from 1955 to 1972

**Algorithms**

Naïve Bayes

AdaBoostM1

DecisionTable

RandomForest

**Attribute Selection**

Correlation

GainRatio

InfoGain

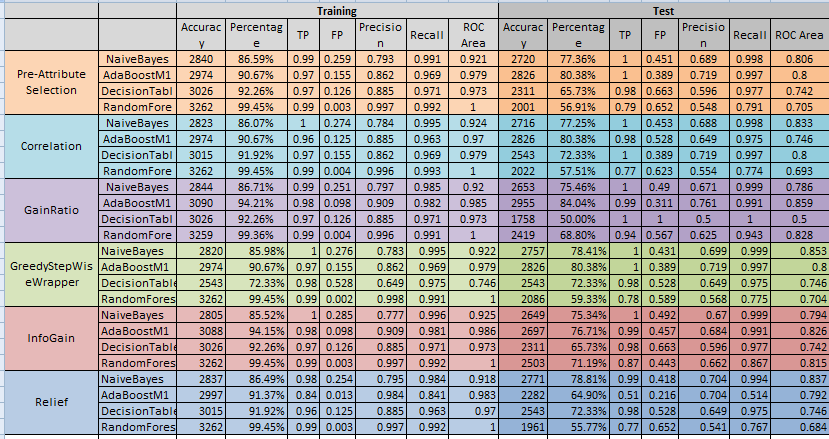
Wrapper - GreedyStepWise

ReliefAttributeEval

# Part 5: Performance Evaluation

The models were compared for performance. Here are the steps to reproduce the performance matrices

1. Run models without any attribute filters:
   1. Open the file titled pre (‘Pre’ meaning pre-attribute selection)
   2. Load CancerData7.arff in Weka, load the 4 model files included in the ‘Pre’ folder.
   3. Load CancerData8.arff file from the ‘Pre’ folder into weka as test dataset.
   4. Run Training and Test datasets with the four loaded models. The results should confirm to the numbers in the worksheet included here.
2. Run models with attribute filters:
   1. There is a folder named ‘Post’ (Post-attribute selection data is included here). There are five sub folders here, each representing both training and test .arff files that were the output from attribute selections.
   2. Open each sub folder, load the train .arff file. A file with number 7 included is the train data, and number 8 is the test data.
   3. Run Training and Test datasets with the four loaded models. The results should confirm to the numbers in the worksheet included here.



# Part 6: Conclusion

* 1. AdaBoostM1, with Relief and InfoGain attribute Selectors give the best FP and ROC.
  2. The models can classify Lung Cancer and Leukemia with a reasonable level of accuracy.
  3. This is because the data set provides age at death. There are 26 age groups and usually Lung Cancer affects mostly people who are old enough to start smoking and be susceptible to other carcinogenic substances.
  4. The objective was to see if these classification algorithms could detect the patterns which are understood when looking at cause of death and age at death. The algorithms clearly could predict with a reasonable degree of accuracy. What is left to be done is to independently compare Lung Cancer and Leukemia against other types of cancers.
  5. The facts revealed in a simple data mining exercise are not accepted by tobacco companies. Please read [this](http://www.who.int/tobacco/media/en/TobaccoExplained.pdf) WHO document.