CSE634 DATA MINING PROJECT

Bakary Classification of Data

using Weka (Version 3.6.10)

Sindhuri Mamidi (109596303)

Sreevathsan Ravichandran (109596710)

Varsha Mohan Paidi (107677677)

1. SELECTION:

1.1 File Preparation

The additional empty sheets in the workbook was deleted and Converted the .xls format into .csv format. To load data into WEKA, we have to put it into a format that will be understood, so we have changed it to .csv format.
Given classes are:
C1: R. Carbonatees AND R. Carbonatees impures
C2: Pyrate
C3: Charcopyrite
C4: Galene
C5: Spahlerite
C6: Sediments terrigenes
When the file loaded and the class attribute was selected as Type de roche, the tool throws error as two attributes have the same name. It was observed that the class attribute was duplicated at the last column and hence the last column(AY) deleted. The tool now allowed to choose the class attribute without any problem.

1.2 Class Verification

Selected the column TYPE DE ROCHE (RockType) as the CLASS attribute.
9 classes were observed in the given data set. The classes 'Type de roche' and ppms (parties par millions) are irrelevant hence removed.

Motivation

The uniqueness and clarity of the attributes and valid data source needed for the KDD process is the motivation for the above step.

Results

The expected input file that could be loaded in to the weka tool for the data mining is now prepared and the ambiguity in the classes and attribute names are handled.

---

2. DATA CLEANING:

2.1 Attribute values correction

1.Since we are interested in only 6 classes, we have merged R. Carbonatees and impures R.Carbonatees into a single class C1.
2.Classes 'Pyyrites' and 'Pyrite' are misspelled. Both of them are merged and name into **pyrate**.
3.'Chalcopyrites' is misspelled leading to 7 classes. It is renamed to **charcopyrite**
4.Similarly 'GalSne' is renamed to **Galene**
5.'S,diments terrigsnes' is renamed to **Sediments terrigenes.**
6. Removed the the second row in the dataset which is empty as the WEKA tool considers an empty row as another instance/record.

While analysing the data set using the WEKA tool,we realised that the dataset for ‘Li’ attribute has divergent values from the rest of the attributes. This is because of the value ‘< 0.3’ in its column while we need only numerical values and no special characters such as ‘<’. Replace this value with number less that 0.3 that is 0.29.

3.DATA PREPARATION

3.1 Principal Component analysis:

Performs a principal components analysis and transformation of the data. Use in conjunction with a Ranker search. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data---default 0.95 (95%). Attribute noise can be filtered by transforming to the PC space, eliminating some of the worst eigenvectors, and then transforming back to the original space.

3.2 Outlier detection:

The criteria used to detect the outlier is the standard deviation, if the value is too high then the those values are replaced by medians. The Null values are also replaced by medians.

3.3 Data Discretization:

My Data1: Equal width Binning is chosen as the discretization method
My Data2: Equal frequency Binning is chosen as the discretization method

---

4. Learning Proper:

The training data and testing data is splitted from the input dataset in the ratio of 1:10 using the Cross-validation functionality provided by the tool, with the folds parameter set to 10. J48 classifier is selected and the classifier is built using the training data. The binarySpilits parameter of the tree in the tool is set to true. This step needs to be executed for the My Data1 and My Data2. Thus, we obtain two classifiers with some predictive accuracy.

Experiments

All the experiments are carried out using preprocessed input dataset generated from the above discussed methods.

Experiment 1

Using all records to find rules for full classification i.e rules describing all classes C1 - C6 simultaneously. The tree and associated metrics are available in the screenshot document.

1.1 My Data1 - Equal Width Binning

Predictive Accuracy: 84.69%

Rules (Classifier)

Rule 1: If the K2O is less than 0.965 and Fe2O3 is between 12.423 and 16.554 then the type of rock is Charcopyrite.
Rule 2: If the K2O is less than 0.965 nad Fe2O3 is not between 12.423 and 16.554 and Zn is less than 3197.8 then the type of rock is R.Carbonatees AND R.Carbonatees impures.
Rule 3: If the K2O is less than 0.965 nad Fe2O3 is not between 12.423 and 16.554 and Zn is greater than 3197.8 then the type of rock is Galene.
Rule 4: If the K2O is greater than 0.965 then the type of rock is Sediments terrigenes.

1.2 My Data2 - Equal Frequency Binning

Predictive Accuracy: 82.65%

Rules (Classifier)

Rule 1: If Al2O3 is greater than 1.67 then the type of rock is Sediments terrigenes.
Rule 2: If Al2O3 is less than 1.67 and S is greater than 9445.5 and Zn is greater than 133.5 then the type of rock is Galene.
Rule 3: If Al2O3 is less than 1.67 and S is greater than 9445.5 and Zn is less than 133.5 then the type of rock is Pyrate.
Rule 4: If Al2O3 is less than 1.67 and S is less than 9445.5 and Zn is greater than 133.5 and MgO is between 8.24 and 16.525 then the type of rock is R. Carbonatees AND R. Carbonatees impures.
Rule 5: If Al2O3 is less than 1.67 and S is less than 9445.5 and Zn is greater than 133.5 and MgO is not between 8.24 and 16.525 then the type of rock is Spahlerite.
Rule 6: If Al2O3 is less than 1.67 and S is less than 9445.5 and Zn is less than 133.5 then the type of rock is R. Carbonatees AND R. Carbonatees impures.

---

2. Experiment 2

Contrasting Class C1 with all other classes:
Set all the values corresponding to the class C1(R. Carbonatees and R. Carbonatees impures) as ‘one’ and the rest of the class attributes to ‘zero’.

2.1 My Data1 - Equal Width Binning

Predictive Accuracy: 83.67%

Rules (Classifier)

Rule 1: If the K2O is less than 0.965 , Fe2O3 is less than 4.161, Zn is less than 3197.8,MnO is between 0.061 to 0.122 and CaO is between 22.79 and 26.58 then the type of rock is “not R. Carbonatees and R. Carbonatees impures”.
Rule 2: If K2O is less than 0.965 and Fe2O3 is less than 4.161, Zn is less than 3197.8,MnO is between 0.061 to 0.122 and CaO is not between 22.79 and 26.58 then the type of rock is “R. Carbonatees and R. Carbonatees impures”.

Rule 3: If K2O is less than 0.965 and Fe2O3 is less than 4.161, Zn is less than 3197.8, MnO is not between 0.061 to 0.122 then the type of rock is "R. Carbonatees and R. Carbonatees impures"
Rule 4: If K2O is less than 0.965 and Fe2O3 is less than 4.161, Zn is more than 3197.8 , then the type of rock is "not a R. Carbonatees and R. Carbonatees impures".
Rule 5: If K2O is less than 0.965 and Fe2O3 is less than 4.161, then the type of rock is "not R. Carbonatees and R. Carbonatees impures."
Rule 6: If K2O is more than 0.965 then the type of rock is "not R. Carbonatees and R. Carbonatees impures."

2.2 My Data2 - Equal Frequency Binning

Predictive Accuracy: 94.89%

Rules (Classifier)

Rule 1: If CaO+MgO is less than 28.27 then the type of rock is "not a R. Carbonatees and R. Carbonatees impures."
Rule 2: If CaO+MgO is not less than 28.27 and S is more than 9445.5 then the type of rock is "not a R. Carbonatees and R. Carbonatees impures".
Rule 3: If CaO+MgO is not less than 28.27 and S is less than 9445.5 and Zn is more than 133.5 and Mgo is between 8.24 and 16.52 then the type of rock is a "R. Carbonatees and R. Carbonatees impures".
Rule 4: If CaO+MgO is not less than 28.27 and S is less than 9445.5, Zn is more than 133.5 and Mgo is not in between 8.24 and 16.52 then the type of rock is "not a R. Carbonatees and R. Carbonatees impures."
Rule 5: If CaO+MgO is not less than 28.27 and S is not less than 9445.5, Zn is less than 133.5 and K20 is between 0.385 and 1.91 and S is between 802 and 1280.5 then then the type of rock is "R. Carbonatees and R. Carbonatees impures."
Rule 6: If CaO+MgO is not less than 28.27 and S is not less than 9445.5, Zn is less than 133.5 and K20 is not between 0.385 and 1.91 then then the type of rock is "R. Carbonatees and R. Carbonatees impures."

Experiment 3

The data containing only the attributes suggested by experts are loaded into the tool.

3.1 Non-Contrast learning

3.1.1 My Data1 - Equal Width Binning

Predictive Accuracy: 89.8%

Rules (Classifier)

Rule 1: If S is less than or equal to 1884.0 and Zn is less than or equal to 188.0 and CaO is less than or equal to 13.61, then the rock is of type "Sediments terrigenes"
Rule 2: If S is less than or equal to 1884.0 and Zn is less than or equal to 188.0 and CaO is more than 13.61, then the rock is of type "R. Carbonatees AND R. Carbonatees impures"

Rule 3: If S is less than or equal to 1884.0 and Zn is more than 188.0, then the rock is of type "Spahlerite"
Rule 4: If S is more than 1884.0 and CaO+MgO less than or equal to 3.63, the rock is of type "Sediments terrigenes"
Rule 5: If S is more than 1884.0 and CaO+MgO more than 3.63, Pb less than 113.0 and CaO+MgO less than or equal to 31.48, then the rock is of type "Charcopyrite"
Rule 6: If S is more than 1884.0 and CaO+MgO more than 3.63, Pb less than 113.0 and CaO+MgO more than 31.48 then the rock is of type "Pyrate"
Rule 7: If S is more than 1884.0 and CaO+MgO more than 3.63, Pb more than 113.0 then the rock is of type "Galene"

3.1.2 My Data2 - Equal Frequency Binning

Predictive Accuracy: 86.7%

Rules (Classifier)

Rule 1: If CaO+MgO is less than 28.275, then the rock is of type "Sediments terrigenes"
Rule 2: If CaO+MgO is more than 28.275 and Pb is more than 5694.5, then the rock is of type Galene
Rule 3: If CaO+MgO is more than 28.275 and Pb is less than 5694.5 and S is more than 9445.5, then the rock is of type Pyrate
Rule 4: If CaO+MgO is more than 28.275 and Pb is less than 5694.5 and S is less than 9445.5, and Zn is more than 133 and MgO is between 8.24 to 16.525 then the rock is of type "R. Carbonatees AND R. Carbonatees impures"
Rule 5: If CaO+MgO is more than 28.275 and Pb is less than 5694.5 and S is less than 9445.5, and Zn is more than 133 and Mg is not between 8.24 to 16.525 then the rock is of type "Spahlerite"

3.2 Contrast Learning

3.2.1 My Data1 - Equal Width Binning

Predictive Accuracy: 89.8%

Rules (Classifier)

Rule 1: If S is more than 1884, then the rock is **not** of type "R. Carbonatees AND R. Carbonatees impures"
Rules 2: If S is less than 1884 and Fe2O3 is less than 0.32 then the rock is of type "R. Carbonatees AND R. Carbonatees impures"
Rule 3: If S is less than 1884 and Fe2O3 is more than 0.32 and if CaO+MgO less than or equal to 26.67, then the rock is **not** of type "R. Carbonatees AND R. Carbonatees impures"
Rule 4: If S is less than 1884 and Fe2O3 is more than 0.32 and if CaO+MgO more than 26.67 and Zn is less than or equal to 188, then the the rock is of type "R. Carbonatees AND R. Carbonatees impures"
Rule 5: If S is less than 1884 and Fe2O3 is more than 0.32 and if CaO+MgO more than 26.67 and Zn is more than 188, then the rock is **not** of type "R. Carbonatees AND R. Carbonatees impures"

3.2.2 My Data2 - Equal Frequency Binning

Predictive Accuracy: 94.9%

Rules (Classifier)

Rule 1: If the CaO+MgO is more than 28.275 and S is more than 9445.5, then the rock is of Type "R. Carbonatees AND R. Carbonatees impures"

Rule 2: If the CaO+MgO is more than 28.275 and S is more than 9445.5 and Zn is more than 133.5 and MgO is between 8.24 to 16.525, Then the Rock is of Type "R. Carbonatees AND R. Carbonatees impures"

Rule 3: If the CaO+MgO is more than 28.275 and S is more than 9445.5 and Zn is more than 133.5 and MgO is not between 8.24 to 16.525, Then the Rock is **not** of Type "R. Carbonatees AND R. Carbonatees impures"

Rule 4: If the CaO+MgO is more than 28.275 and S is more than 9445.5 and Zn is not more than 133.5, then Rock is of Type "R. Carbonatees AND R. Carbonatees impures"

Conclusion:

The KDD process in data mining is carried out successfully with the help of Weka tool. Total of six classifier are presented as part of results and the predictive accuracy is also computed. The suitable classifier could be used to predict the type of rock for the new data.