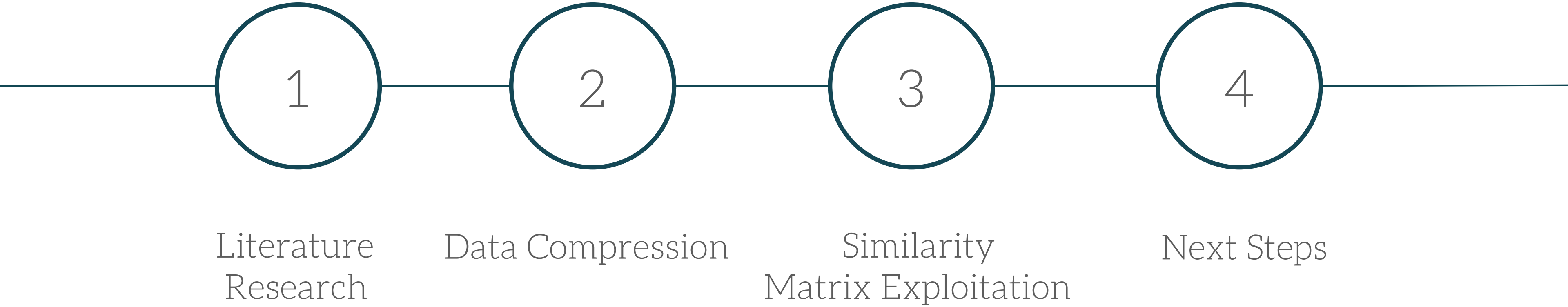
The Mushroom Database

Descriptive Mining II

5. November. 2014

ROSTLAB.

Agenda



1

Literature Research

2

Data Compression

3

Similarity Matrix Exploitation

4

Next Steps

Literature Research

Boriah et. al. [1] compared different similarity methods for nominal data.

They classified them into 3 categories depending on how the similarity matrix is formed:

1. Methods which fill diagonal entries only
2. Methods which fill off-diagonal entries only
3. Methods which fill both diagonal and off-diagonal entries

Literature Research

1. ***Lin:***
   - Use weights for different attributes
   - Higher wieghts for attribute matches which occur frequently in the data set
   - Lower weights to attribute mis-matches occurring infrequently

2. ***Smirnov:***
   - Considers the frequency of other sub-attributes for a given main attribute
   - Higher similarity score when matching attribute frequency is low, and other values are frequent

3. ***Anderberg:***

   Attribute value matches which are infrequent are given a high weightage
   Assigns higher similarity to attribute mismatches which are rare

Data Compression

Deleting data

Delete Redundant Features

Threshold >95%

- Gill Attachment (97.42%)
- Veil Type (100%)
- Veil Color (97.54%)

Delete Redundant Characteristics

Threshold >95%

- Gill Spacing: characteristic 'distant'
- Stalk Root: characteristics 'cup' and 'rhizomorphs'.
- Ring Type: characteristics 'cobwebby', 'sheathing' and 'zone'

Delete Strong Correlated Features

Threshold >95%

- Stalk Surface Above Ring + Stalk Surface Below Ring: 77.01%
- Stalk Color Above Ring + Stalk Color Below Ring: 62.38%

Data Compression

grouping data

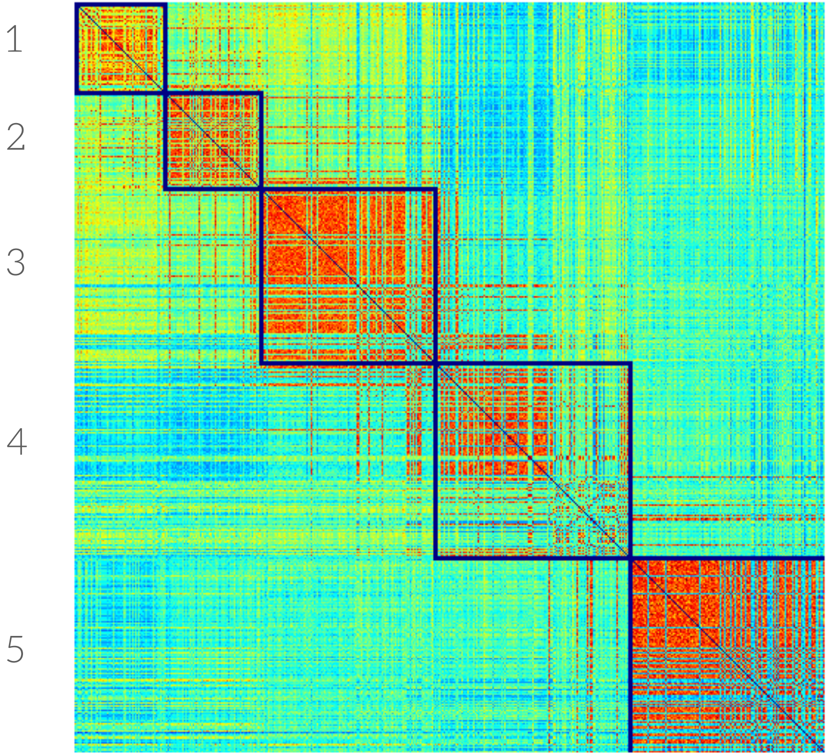
Group Characteristics

Threshold <1%

- Cap Color: 'brown' + 'cinnamon' = 'brown'. 'green' + 'pink' = 'misc'
- Ring Type: 'flaring' + 'none' = 'misc‘
- Odor: 'musty' + 'foul' = 'foul'

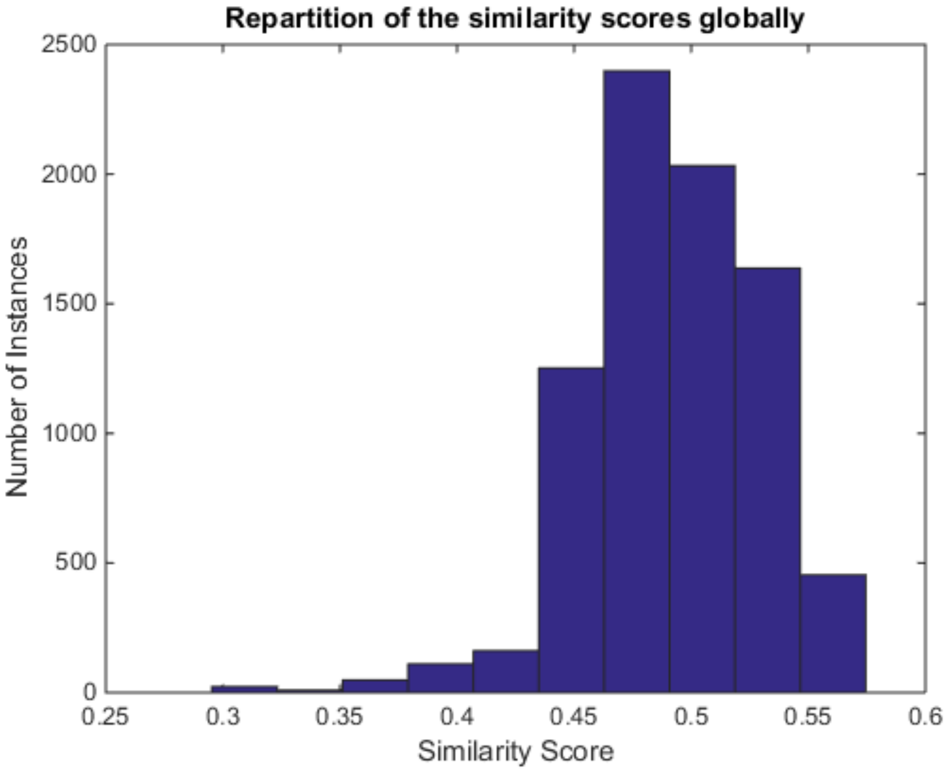
Similarity Matrix Exploitation

The Similarity Matrix is arbitrarily split in 5 groups.



Average Similarity Distribution

Whole Similarity Matrix

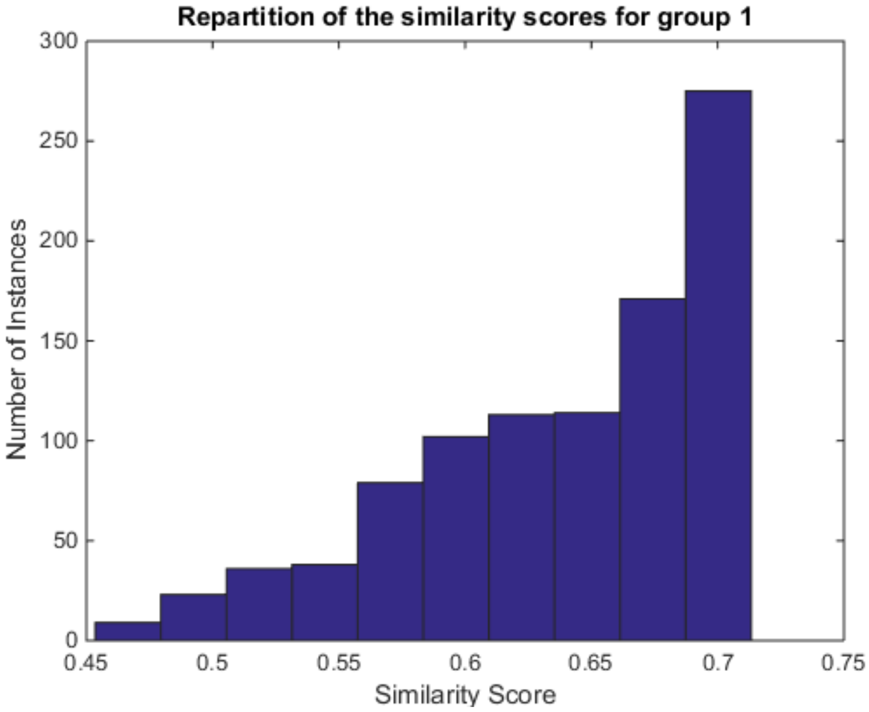


- 0 no attribute in common
- 1 all attributes in common
- For *each* instance with *all other* instances

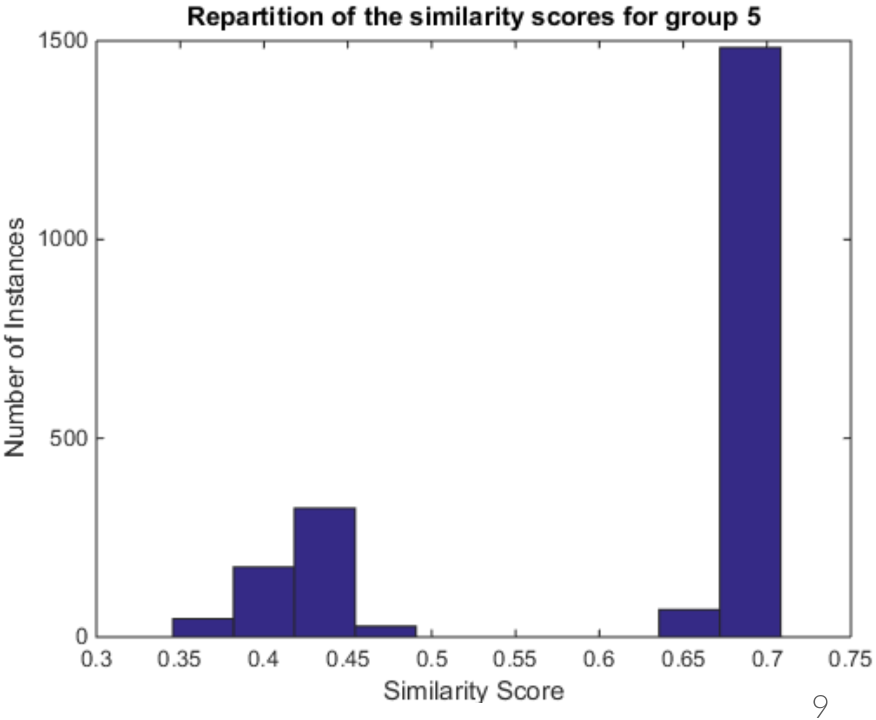
→ Helps to identify outliers

Average Similarity Distribution

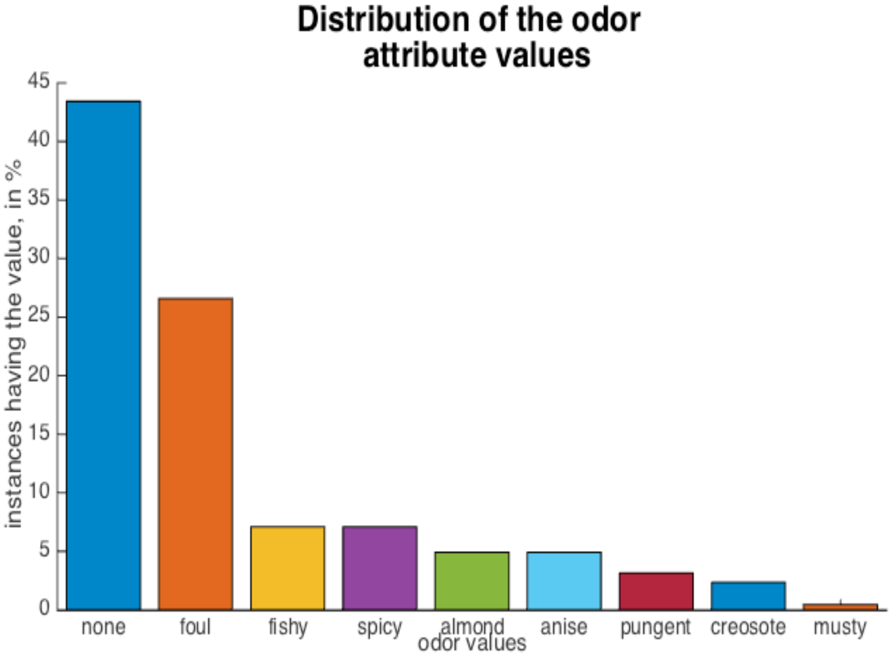
Group 1

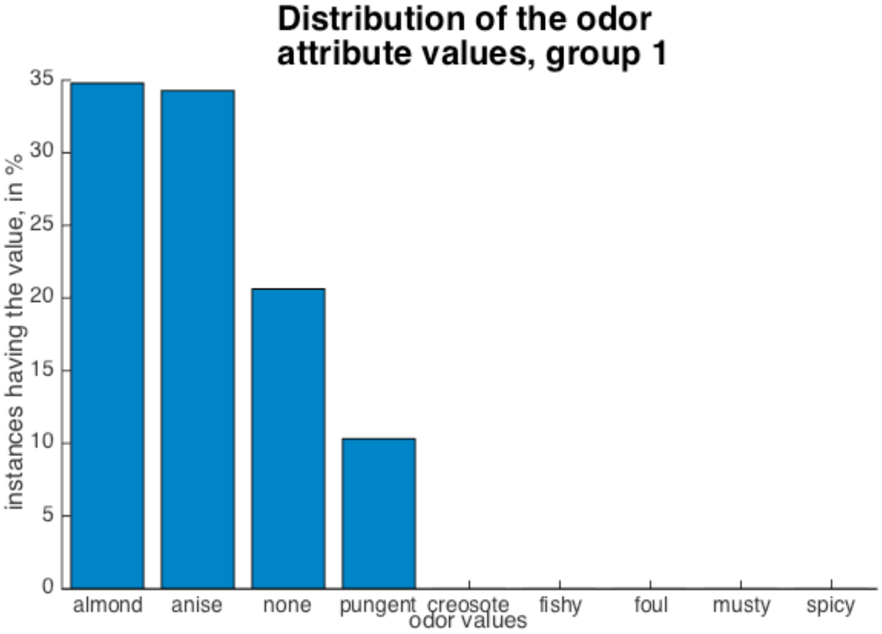


Group 5



Attribute Value Distibution





→External Link to the Wiki

Next Steps

- Further study of the naïve, not compressed Similarity matrix
- Comparative study of its compressed version
- Refine the grouping using Weka
- See how we can apply the Frequent Item Dataset method
- Test some of the new similarity measures, when applicable