Sharing data using multiple imputation

Sicco Verwer, Susan van den Braak, Sunil Choenni

Research and Documentation Centre (WODC)
Dutch Ministry of Security and Justice

Slide nr 1

- Motivation
- Multiple imputation
  - Initial results
- Imputing network structures
  - Initial results
- Future work

Chamber of commerce data

- Propietary data containing
  - Legal persons including:
    - Name
    - Adress
  - Natural persons including:
    - Name
    - Adress
    - Date of birth
  - Functions:
    - Functional description
    - Legal/natural person performing the function
    - Legal person for which the function is performed

Coc data at the WODC

- We are not allowed access to the CoC data
- Copying the data is prohibited
- In order to test new methods, we require a realistic artificial copy of the CoC data
  - Similarly distributed legal persons
  - Similarly distributed natural persons
  - Similar links (functions) between persons

  - Impossible to draw real conclusions on an individual level from the fake copy
  - Possible to test real methods on the fake copy

Creating a realistic copy

- Generate a completely new dataset by imputing it based on existing data:

value	maint.	doors	persons	boot	safety	class
vhigh	low	4	4	small	med	unacc
high	low	5more	2	small	med	unacc
med	high	3	2	big	high	unacc
low	med	4	more	med	high	vgood
vhigh	high	4	4	med	high	unacc

Add empty rows

value	maint.	doors	persons	boot	safety	class
vhigh	low	4	4	small	med	unacc
high	low	5more	2	small	med	unacc
med	high	3	2	big	high	unacc
low	med	4	more	med	high	vgood
vhigh	high	4	4	med	high	unacc
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?

Impute their values

value	maint.	doors	persons	boot	safety	class
vhigh	low	4	4	small	med	unacc
high	low	5more	2	small	med	unacc
med	high	3	2	big	high	unacc
low	med	4	more	med	high	vgood
vhigh	high	4	4	med	high	unacc
high	med	4	more	big	high	acc
vhigh	med	2	2	big	low	unacc
vhigh	med	4	4	big	low	unacc
high	low	2	2	small	low	unacc
high	low	2	more	small	high	unacc
high	vhigh	3	2	med	med	unacc

Share the imputed values

value	maint.	doors	persons	boot	safety	class
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
high	med	4	more	big	high	acc
vhigh	med	2	2	big	low	unacc
vhigh	med	4	4	big	low	unacc
high	low	2	2	small	low	unacc
high	low	2	more	small	high	unacc
high	vhigh	3	2	med	med	unacc

Multiple imputation

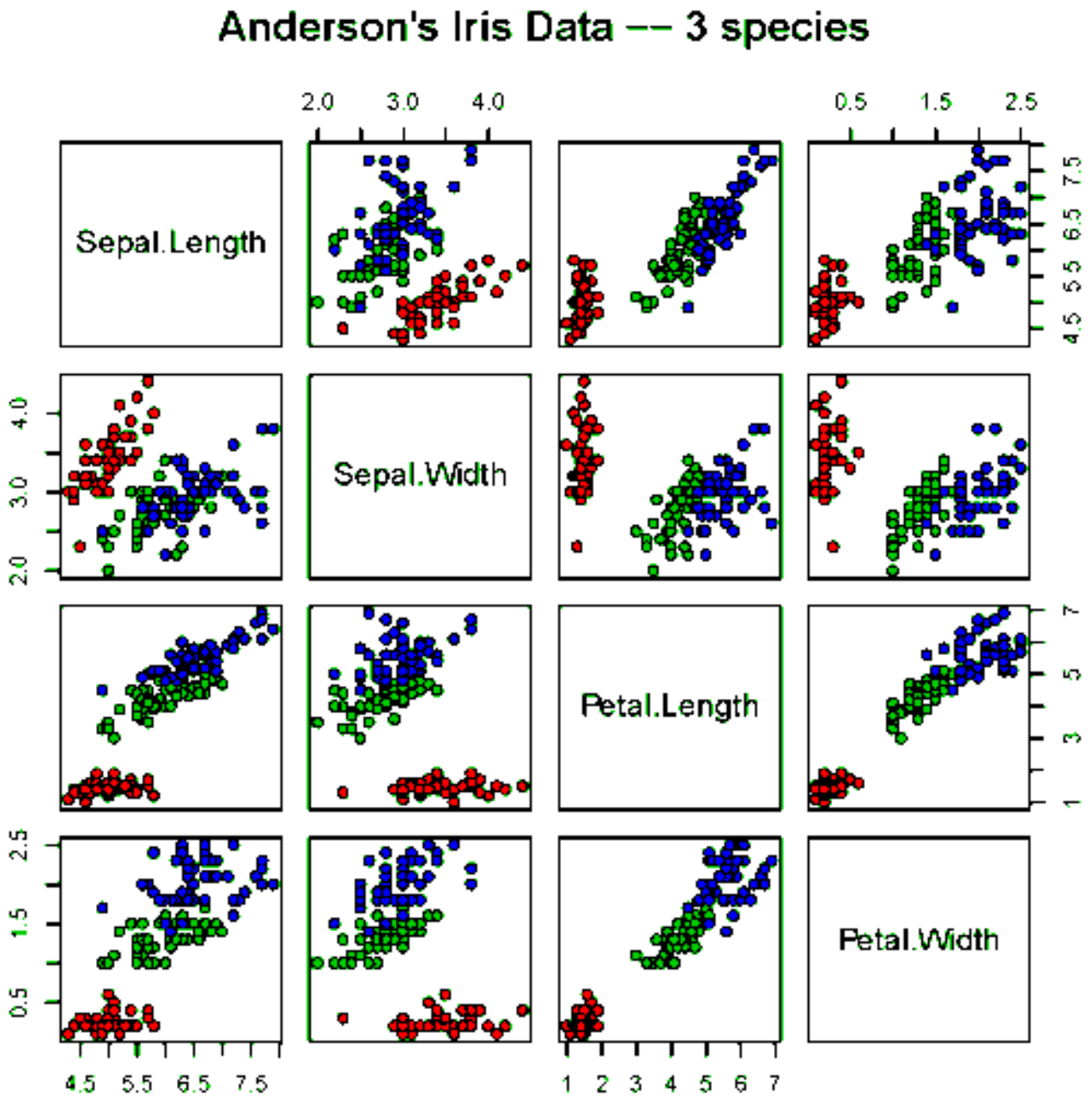
- Technique for algorithms that require complete datasets
- Create multiple (typically 5) imputations of all missing values
- Run an analysis algorithm om all imputations
- Average the results to obtain an estimate of the result on the original (non-imputed) data

Results

- We generate artificial datastes using MICE (Multiple Imputation using Chained Equations, a tool in R) based on 4 datasets from the UCI repository: iris, wine, car, and breast cancer
- We test:
  - The difficulty of distinguishing the original from the generated data using three standard classifiers: naive Bayes, decision tree, and nearest neighbor
  - The difference in classifier performance on the original and generated datasets

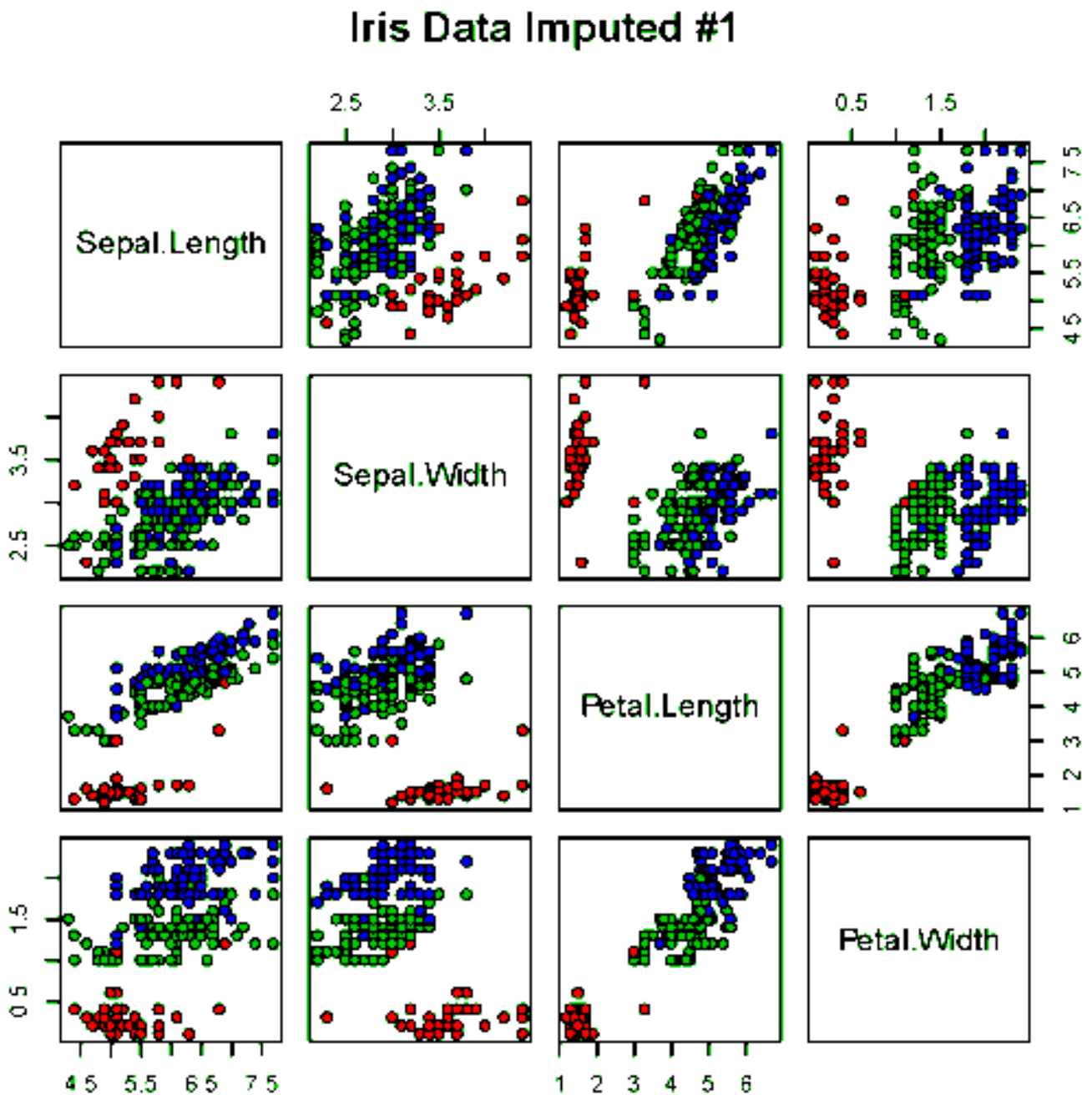
Results

150 instances



Results

1500 instances



Results: naive Bayes

	1	2	3	4	5	Average
Iris	0,52	0,52	0,51	0,55	0,56	0,532
Wine	0,55	0,53	0,56	0,57	0,53	0,548
Car	0,51	0,52	0,51	0,53	0,51	0,516
Breast cancer	0,52	0,5	0,5	0,52	0,5	0,508

Results: decision tree

	1	2	3	4	5	Average
Iris	0,48	0,54	0,5	0,48	0,52	0,504
Wine	0,57	0,57	0,67	0,7	0,61	0,624
Car	0,53	0,54	0,55	0,55	0,54	0,542
Breast cancer	0,55	0,57	0,63	0,57	0,54	0,572

Results: nearest neighbor

	1	2	3	4	5	Average
Iris	0,54	0,55	0,55	0,49	0,61	0,548
Wine	0,54	0,54	0,53	0,56	0,49	0,532
Car	NA	NA	NA	NA	NA	NA
Breast cancer	0,58	0,6	0,58	0,57	0,57	0,58

Results: classification

	Iris		Wine		Car		Breast cancer	
	Avg	Orig	Avg	Orig	Avg	Orig	Avg	Orig
Naive Bayes	0,906	0,96	0,952	0,98	0,862	0,86	0,932	0,96
Decision tree	0,926	0,93	0,884	0,9	0,862	0,94	0,93	0,94
Nearest neighbor	0,9	0,96	0,722	0,75	NA	NA	0,93	0,95

Car data
naive Bayes class distribution

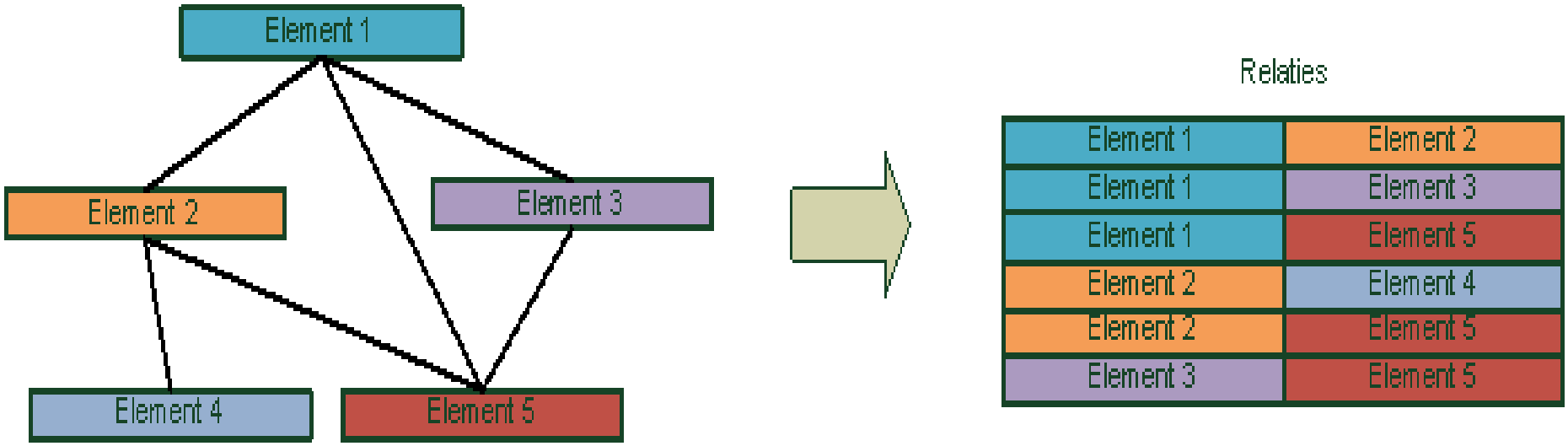
	Generated			Original		
	low	med	high	low	med	high
unacc	0.416	0.288	0.30	0.48	0.30	0.229
acc	0.089	0.428	0.483	0.00	0.469	0.531
good	0.174	0.413	0.413	0.00	0.565	0.435
vgood	0.747	0.107	0.147	0.00	0.00	1.00

Slide nr 1

- Motivation
- Multiple imputation
  - Initial results
- Imputing network structures
  - Initial results
- Future work

Imputing network structure

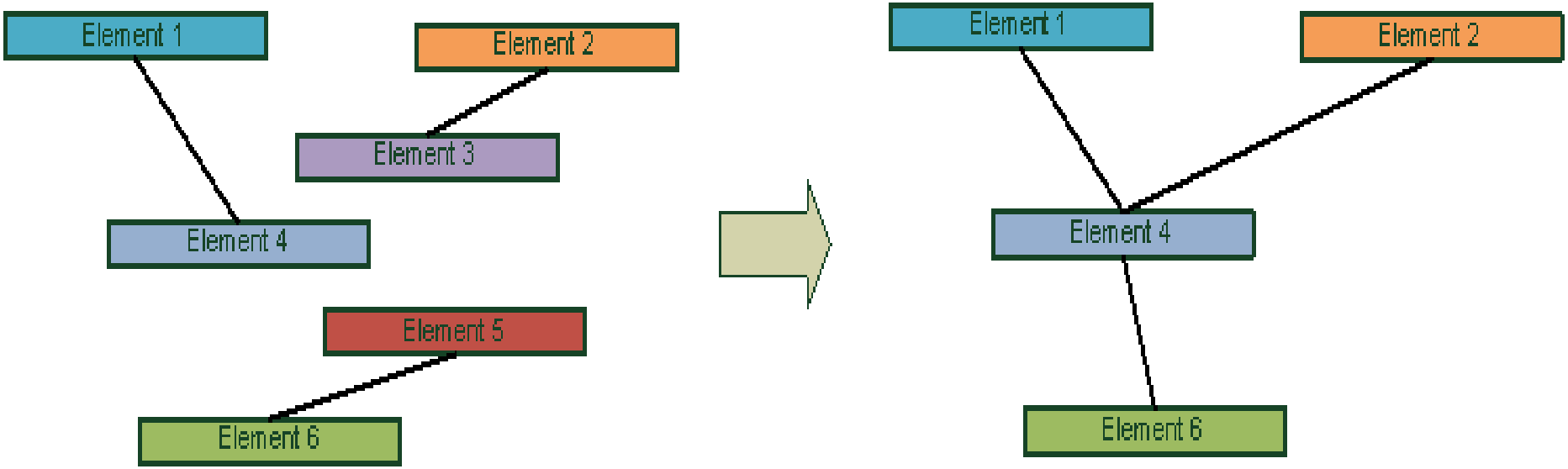
- Impute links instead of elements:



- Include network attributes such as number of incomming and outgoing links, length of path to top and bottom, etc.

Imputing network structure

- Combine generated elements based on their similarity:



Imputing network structure

- Persons in the CoC data are combined using clustering and matching techniques:
  - Cluster legal persons, the cluster sizes are termined by the average number of incomming/outgoing links
  - Match the links of natural persons to the generated structure of legal persons based on their path lengths

- The result is a network structure!
- But is it realistic?

Results before clustering

Classified as	True value	
	Original	Artificial
Original legal	10926	4074
Artificial legal	5385	9615
Original natural	9408	5592
Artificial natural	2125	12875

Results after clustering

Classified as	True value	
	Original	Artificial
Original legal	13373	4425
Artificial legal	1627	10575
Original natural	13067	4910
Artificial natural	1933	10090

Future work

- Test the generated network structure
  - using network classifiers, or frequent subgraphs
- Adjust imputation methods towards keeping classification results intact
  - accuracy and classifier ordering
- Impute an artificial network structure based on the original network structure
- Instead of generating an entire dataset or database, impute only values necessary for privacy guarantees (i.e. k-anonymity) copy all others