



TIỀN XỬ LÝ DỮ LIỆU 2

Giảng viên: Nguyễn Tu Trung
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2021

Nội dung

- ❖ Chuẩn bị dữ liệu
- ❖ Nạp dữ liệu (Loading the Data)
- ❖ Hiển thị dữ liệu
- ❖ Chuẩn hóa dữ liệu về $[0,1]$
- ❖ Thay thế các giá trị bị thiếu
- ❖ Chuyển đổi: Numeric to Nomial

Chuẩn bị dữ liệu

❖ File dữ liệu:
`inputdata.csv`

❖ Gồm các
trường =>

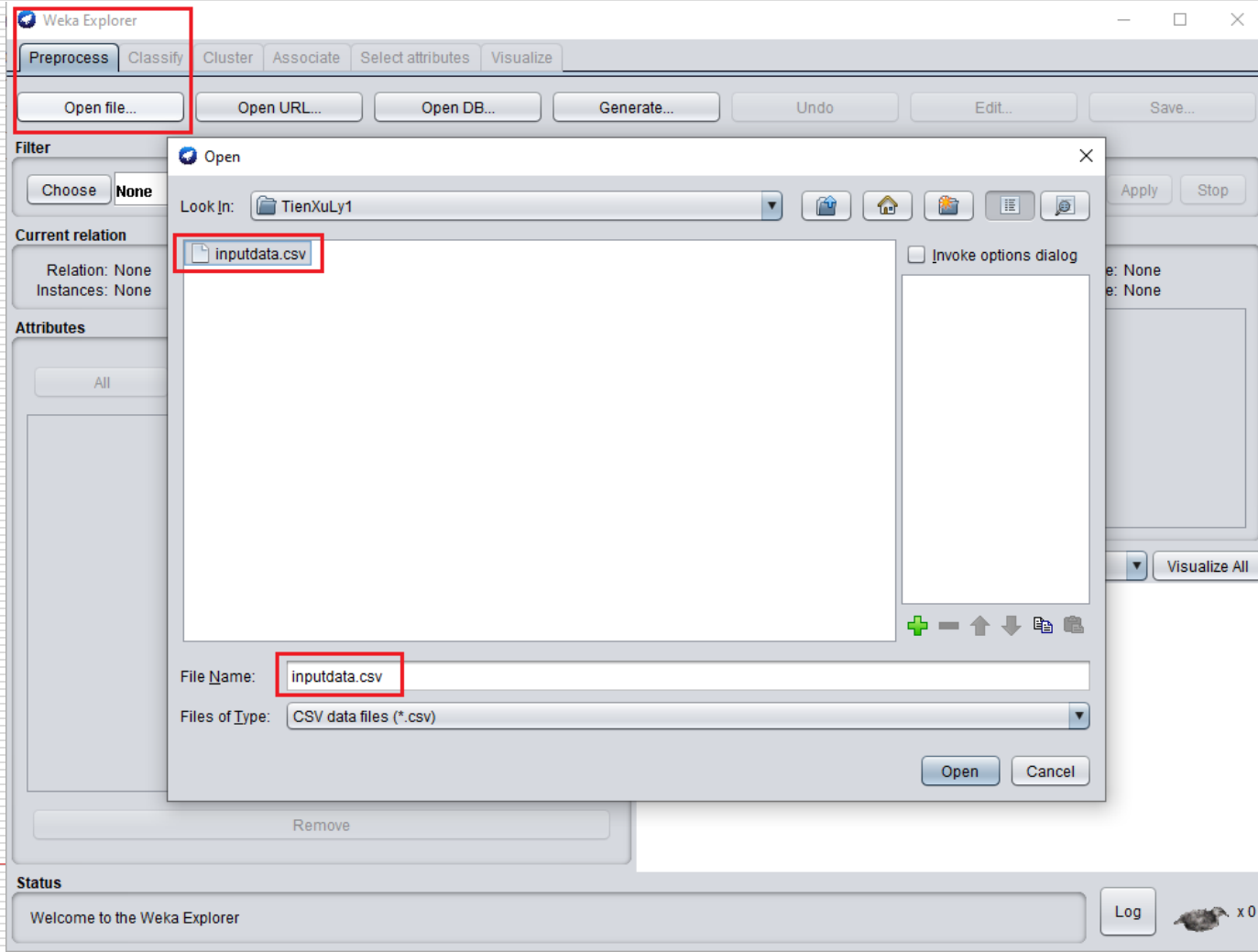
```
Attribute,Type,Mean,StdDev,Missing
surgery,Numeric,1.398,0.49,1 (0%)
Age,Numeric,1.64,2.174,0 (0%)
Hospital Number,Numeric,1085888.833,1529800.909,0 (0%)
rectal temperature,Numeric,38.168,0.732,60 (20%)
pulse,Numeric,71.913,28.631,24 (8%)
respiratory rate,Numeric,30.417,17.642,58 (19%)
temperature of extremities,Numeric,2.348,1.045,56 (19%)
peripheral pulse,Numeric,2.017,1.042,69 (23%)
mucous membranes,Numeric,2.854,1.62,47 (16%)
capillary refill time,Numeric,1.306,0.478,32 (11%)
pain,Numeric,2.951,1.308,55 (18%)
peristalsis,Numeric,2.918,0.977,44 (15%)
abdominal distension,Numeric,2.266,1.065,56 (19%)
nasogastric tube,Numeric,1.755,0.649,104 (35%)
nasogastric reflux,Numeric,1.582,0.805,106 (35%)
nasogastric reflux PH,Numeric,4.708,1.982,247 (82%)
rectal examination,Numeric,2.758,1.251,102 (34%)
abdomen,Numeric,3.692,1.492,118 (39%)
packed cell volume,Numeric,46.295,10.419,29 (10%)
total protein,Numeric,24.457,27.475,33 (11%)
abdominocentesis appearance,Numeric,2.037,0.805,165 (55%)
abdomcentesis total protein,Numeric,3.02,1.969,198 (66%)
```

Chuẩn bị dữ liệu

surgery	Age	Hospital N	rectal tem	pulse	respirator	temperatu	peripheral	mucous m	capillary r	pain	peristalsis	abdominal	nasogastric	nasogastric	nasogastric	rectal exa	abdomen	p
2	1	530101	38.5	66	28	3	3 ?		2	5	4	4 ?	?	?		3	5	
1	1	534817	39.2	88	20 ?		?	4	1	3	4	2 ?	?	?		4	2	
2	1	530334	38.3	40	24	1	1	3	1	3	3	1 ?	?	?		1	1	
1	9	5290409	39.1	164	84	4	1	6	2	2	4	4	1	2	5	3 ?		
2	1	530255	37.3	104	35 ?		?	6	2 ?		?	?	?	?	?	?	?	
2	1	528355	?	?	?	2	1	3	1	2	3	2	2	1 ?		3	3 ?	
1	1	526802	37.9	48	16	1	1	1	1	3	3	3	1	1 ?		3	5	
1	1	529607	?	60 ?		3 ?	?		1 ?		4	2	2	1 ?		3	4	
2	1	530051	?	80	36	3	4	3	1	4	4	4	2	1 ?		3	5	
2	9	5299629	38.3	90 ?		1 ?		1	1	5	3	1	2	1 ?		3 ?		
1	1	528548	38.1	66	12	3	3	5	1	3	3	1	2	1	3	2	5	
2	1	527927	39.1	72	52	2 ?		2	1	2	1	2	1	1 ?		4	4	
1	1	528031	37.2	42	12	2	1	1	1	3	3	3	3	1 ?		4	5 ?	
2	9	5291329	38	92	28	1	1	2	1	1	3	2	3 ?		7.2	1	1	
1	1	534917	38.2	76	28	3	1	1	1	3	4	1	2	2 ?		4	4	
1	1	530233	37.6	96	48	3	1	4	1	5	3	3	2	3	4.5	4 ?		
1	9	5301219	?	128	36	3	3	4	2	4	4	3	3 ?	?		4	5	
2	1	526639	37.5	48	24 ?		?	?	?	?	?	?	?	?	?	?	?	
1	1	5290481	37.6	64	21	1	1	2	1	2	3	1	1	1 ?		2	5	
2	1	532110	39.4	110	35	4	3	6 ?	?		3	3 ?	?	?	?	?		
1	1	530157	39.9	72	60	1	1	5	2	5	4	4	3	1 ?		4	4	
2	1	529340	38.4	48	16	1 ?		1	1	1	3	1	2	3	5.5	4	3	
1	1	521681	38.6	42	34	2	1	4 ?		2	3	1 ?	?	?		1 ?		
1	9	534998	38.3	130	60 ?		3 ?		1	2	4 ?	?	?	?	?	?	?	
1	1	533692	38.1	60	12	3	3	3	1 ?		4	3	3	2	2 ?	?		
2	1	529518	37.8	60	42 ?		?	?	1 ?	?	?	?	?	?	?	?	?	
1	1	530526	38.3	72	30	4	3	3	2	3	3	3	2	1 ?		3	5	
1	1	528653	37.8	48	12	3	1	1	1 ?		3	2	1	1 ?		1	3	
1	1	5279442	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	

Nạp dữ liệu (Loading the Data)

- ❖ Trong Weka Explorer, chọn tab Preprocess
- ❖ Chọn Open file => Chọn file inputdata.csv



Hiển thị dữ liệu

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose **None** Apply Stop

Current relation
Relation: inputdata Instances: 300 Attributes: 28 Sum of weights: 300

Selected attribute
Name: peripheral pulse Missing: 69 (23%) Distinct: 4 Type: Numeric Unique: 0 (0%)

Attributes
All None Invert Pattern

No.	Name
1	<input type="checkbox"/> surgery
2	<input type="checkbox"/> Age
3	<input type="checkbox"/> Hospital Number
4	<input type="checkbox"/> rectal temperature
5	<input type="checkbox"/> pulse
6	<input type="checkbox"/> respiratory rate
7	<input type="checkbox"/> temperature of extremities
8	<input checked="" type="checkbox"/> peripheral pulse
9	<input type="checkbox"/> mucous membranes
10	<input type="checkbox"/> capillary refill time
11	<input type="checkbox"/> pain
12	<input type="checkbox"/> peristalsis
13	<input type="checkbox"/> abdominal distension
14	<input type="checkbox"/> nasogastric tube
15	<input type="checkbox"/> nasogastric reflux
16	<input type="checkbox"/> nasogastric reflux PH
17	<input type="checkbox"/> rectal examination
18	<input type="checkbox"/> ...

Remove

Statistic	Value
Minimum	1
Maximum	4
Mean	2.017
StdDev	1.042

Class: cp_data (Num) Visualize All

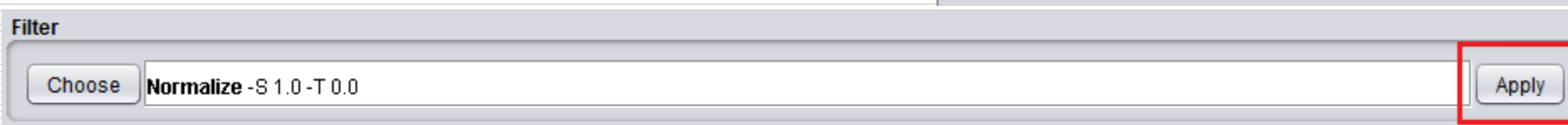
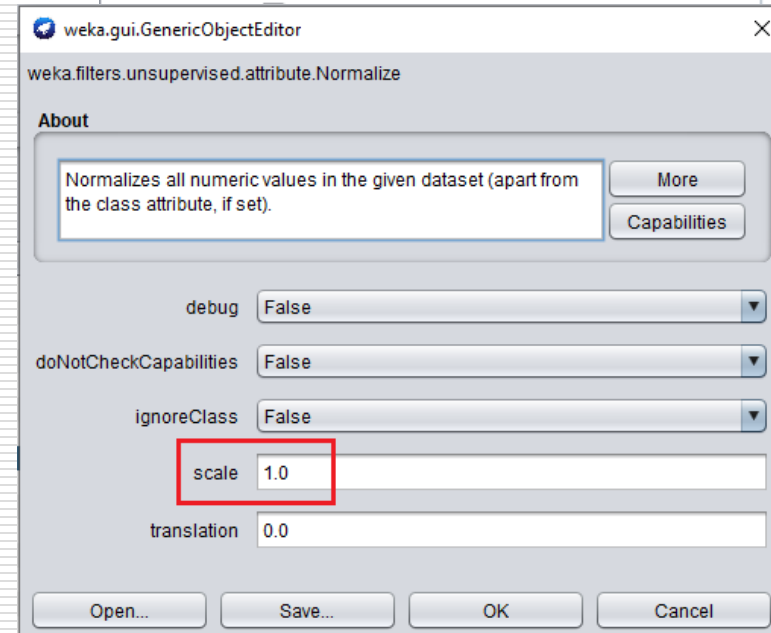
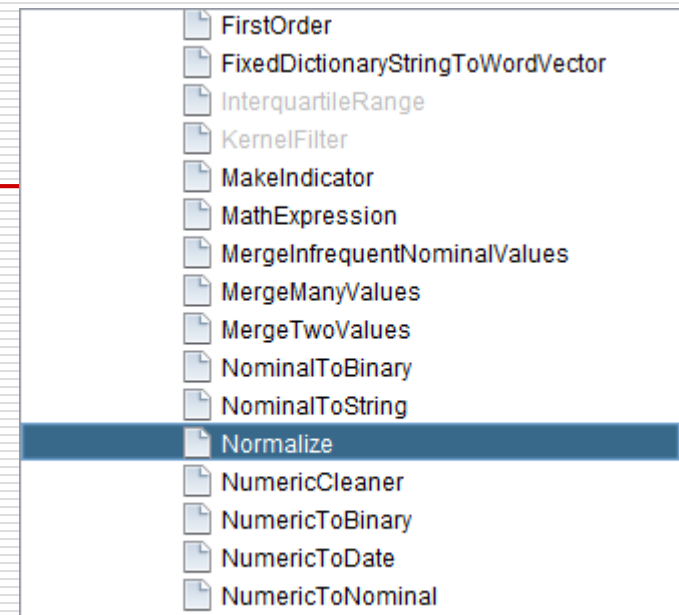
Value	Frequency
1	5
2.5	103
4	8

Hiển thị dữ liệu

- ❖ Nhận xét dữ liệu:
 - ❖ 300 mẫu dữ liệu
 - ❖ 28 thuộc tính dạng số (Numeric)
 - ❖ 30% dữ liệu bị thiếu giá trị
- ❖ => Cần thực hiện các thao tác tiền xử lý

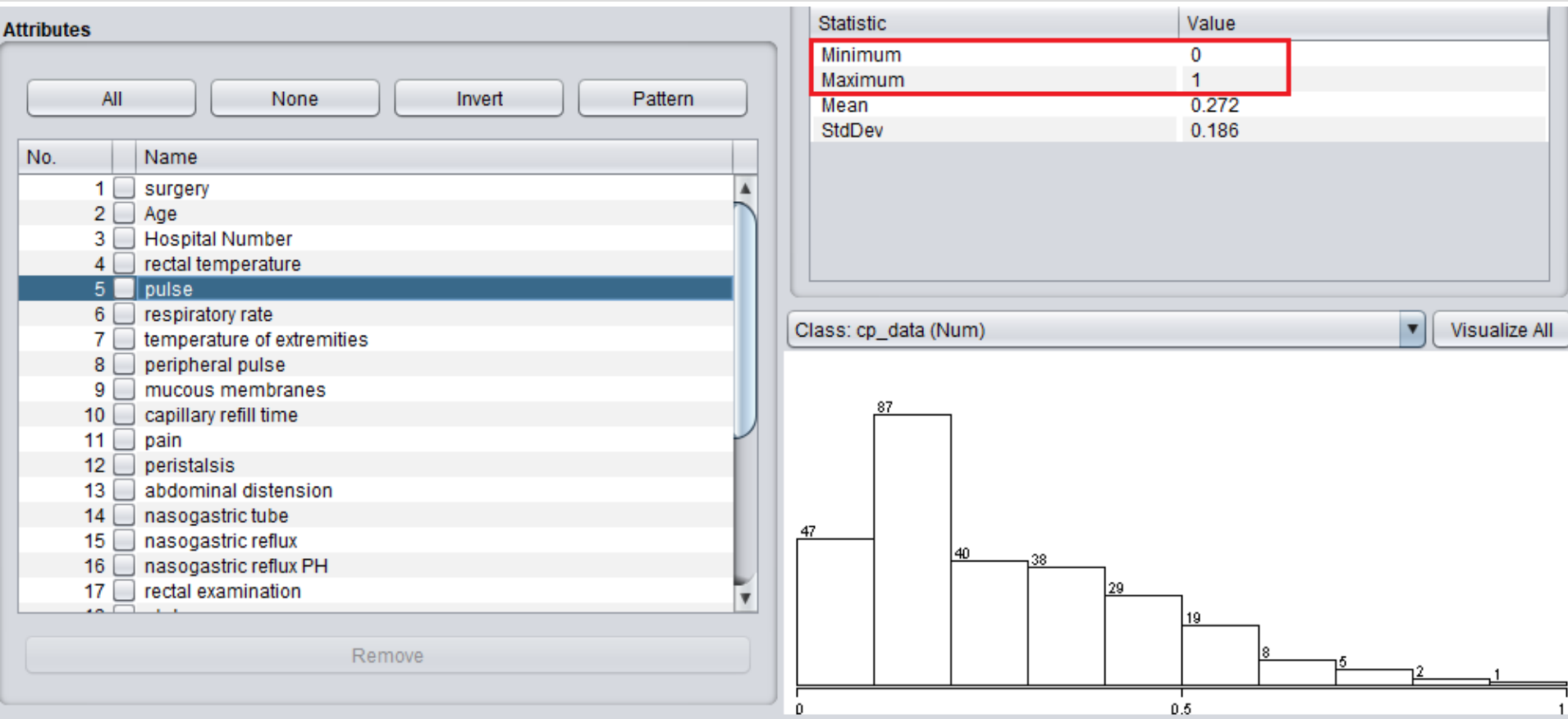
Chuẩn hóa dữ liệu về [0,1]

- ❖ Trong group Filter, nhấn nút “Choose” => *filters* > *unsupervised* > *attribute* > *Normalize*
- ❖ Nhấn chuột trái vào Textbox bên phải nút “Choose”
- ❖ Trong hộp thoại hiện ra
- ❖ Nhập “1” - ứng với chỉ số trường Id vào hộp *attributeIndices*
- ❖ Nhấn OK => Nhấn Apply



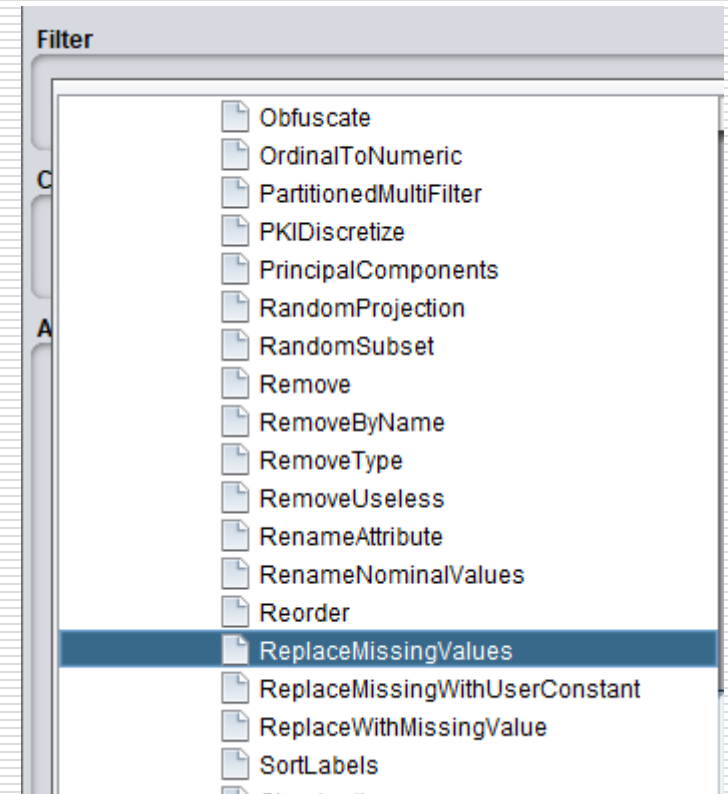
Chuẩn hóa dữ liệu về [0,1]

❖ Sau khi thực hiện:



Thay thế các giá trị bị thiếu

- ❖ Chú ý: Thực hiện trước hoặc sau khi chuẩn hóa tùy theo yêu cầu (có cần chuẩn hóa hay không)
- ❖ Trong group Filter, nhấn nút “Choose” => *filters* > *unsupervised* > *attribute* > *ReplaceMissingValues*
- ❖ Nhấn Apply



Thay thế các giá trị bị thiếu

❖ Sau khi thực hiện: **Thống kê không còn bị thiếu**

Filter

Choose **ReplaceMissingValues** Apply Stop

Current relation

Relation: inputdata-weka.filters.unsupervised.attribute.... Attributes: 28
Instances: 300 Sum of weights: 300

Attributes

All None Invert Pattern


No.	Name
1	<input checked="" type="checkbox"/> surgery
2	<input type="checkbox"/> Age
3	<input type="checkbox"/> Hospital Number
4	<input type="checkbox"/> rectal temperature
5	<input type="checkbox"/> pulse
6	<input type="checkbox"/> respiratory rate
7	<input type="checkbox"/> temperature of extremities
8	<input type="checkbox"/> peripheral pulse
9	<input type="checkbox"/> mucous membranes
10	<input type="checkbox"/> capillary refill time
11	<input type="checkbox"/> pain
12	<input type="checkbox"/> peristalsis
13	<input type="checkbox"/> abdominal distension
14	<input type="checkbox"/> nasogastric tube

Selected attribute

Name: **surgery** Missing: 0 (0%) Distinct: 3 Type: Numeric Unique: 1 (0%)

Statistic	Value
Minimum	1
Maximum	2
Mean	1.398
StdDev	0.489

Class: cp_data (Num) Visualize All



Chuyển đổi: Numeric to Nomial

- ❖ Khác với discretization (rời rạc hóa)
- ❖ Chuyển đổi từ giá trị số thành giá trị Nomial mà không chia thành các giỏ (bin)
- ❖ Có bao nhiêu giá trị số ứng với bấy nhiêu giá trị Nomial
- ❖ Trước khi chuyển:

Current relation

Relation: inputdata
Instances: 300

Attributes: 28
Sum of weights: 300

Attributes

All None Invert Pattern

No.	Name
1	surgery
2	Age
3	Hospital Number
4	rectal temperature
5	pulse
6	respiratory rate
7	temperature of extremities
8	peripheral pulse
9	mucous membranes

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	9
Mean	1.64
StdDev	2.174

Class: cp_data (Num) Visualize All

Chuyển đổi: Numeric to Nomial

- ❖ Trong group Filter, nhấn nút “Choose” => *filters* > *unsupervised* > *attribute* > *ReplaceMissingValues*
- ❖ Nhấn Apply

- MathExpression
- MergeInfrequentNominalValues
- MergeManyValues
- MergeTwoValues
- NominalToBinary
- NominalToString
- Normalize
- NumericCleaner
- NumericToBinary
- NumericToDate
- NumericToNominal**
- NumericTransform

Filter

Choose **NumericToNominal -R first-last** Apply

Current relation Selected attribute

Filter

Choose **NumericToNominal -R first-last** Apply Stop

Current relation

Relation: inputdata-weka.filters.unsupervised.attribute.... Attributes: 28
Instances: 300 Sum of weights: 300

Attributes

All None Invert Pattern

No.	Name
1	Surgery
2	Age
2	Hospital Number
4	rectal temperature
5	pulse

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	1	276	276.0
2	9	24	24.0