Data Mining: Preprocessing

Jef Wijsen

Université de Mons (UMONS)

Outline

The KDD Process

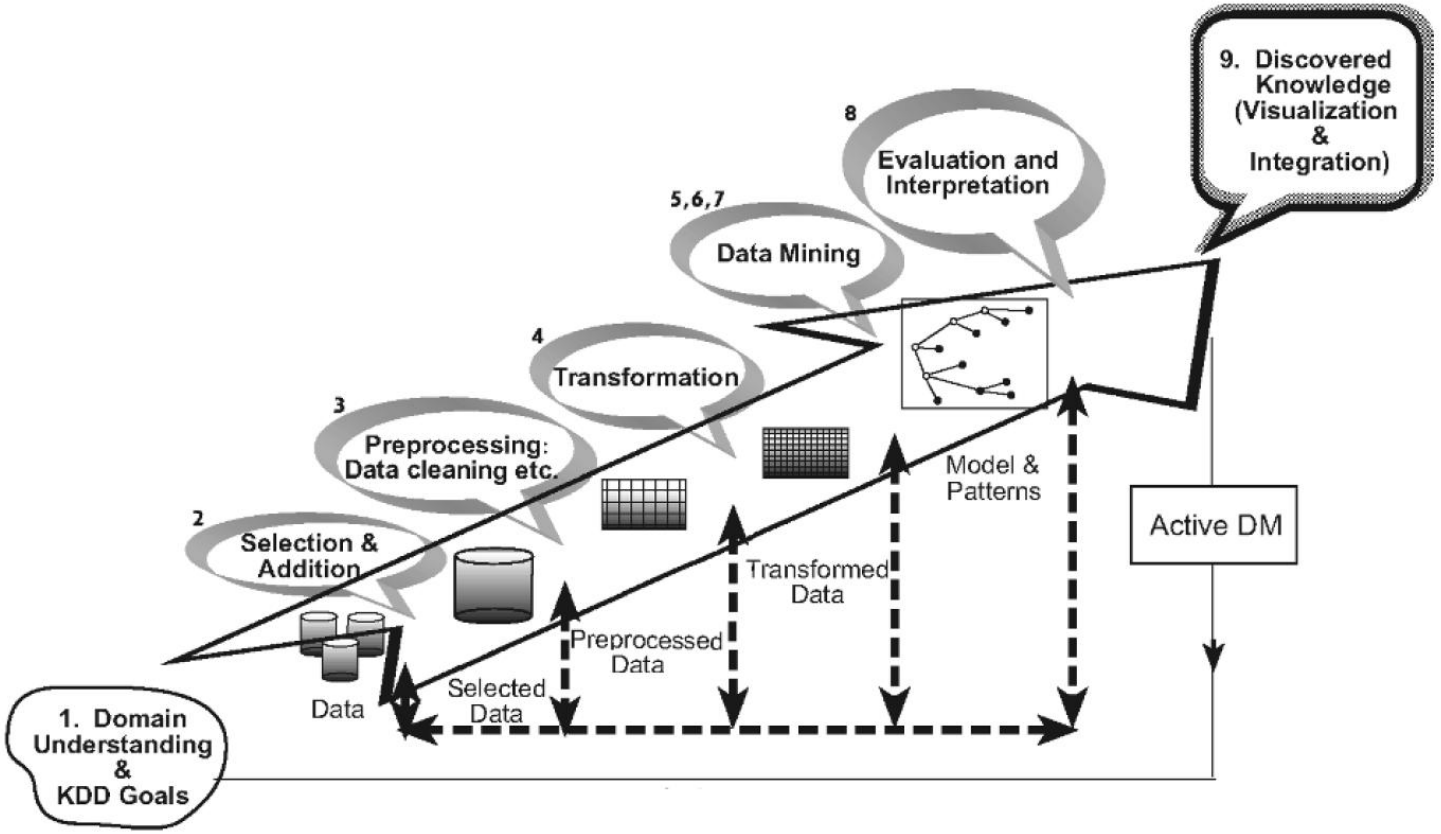


Fig. 1.1. The Process of Knowledge Discovery in Databases.

Source: Oded Maimon, Lior Rokach (Eds.): *The Data Mining and Knowledge Discovery Handbook* (2nd Edition). Springer, 2010

Nettoyage : Exemples

Principe du GIGO (*Garbage In Garbage Out*)...

- Compléter les valeurs manquantes et NULLs.
- Corriger les fautes de frappe et autres erreurs.
- Unifier les synonymes, par ex. 1=M, 2=F.
- Uniformiser les données exprimées en unités différentes, par ex. BEF et EURO.
- ...

Les données étant clairement erronées mais ne pouvant être corrigées sont effacées.

Transformation : Exemples

- Convertir les adresses en coordonnées.
- Calculer la vente quotidienne à partir des ventes individuelles.
- Normaliser les variables entre 0 et 1.
- Remplacer toute valeur x d'un attribut A avec moyenne μ et variance σ^2 par $\frac{x-\mu}{\sigma}$ afin d'arriver à une moyenne de 0 et une variance de 1.
- ...

Outline

Outline

Exercices

17.1 INTRODUCTION TO THE EXPLORER INTERFACE

- Exercises 17.1.1–17.1.7
- The Visualize Panel

Source: Ian H. Witten and Eibe Frank: *Data Mining. Practical Machine Learning Tools and Techniques* (3rd Edition). Morgan Kaufmann, 2011

Outline

Arff

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
% Other attribute types are string and date

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
```

AddExpression

Creates a new attribute by applying a mathematical expression to existing attributes.
Example: `a1^2*a5/log(a7*4.0)`

Center

Centers all numeric attributes to have zero mean (apart from the class attribute, if set).

Discretize

Discretizes a range of numeric attributes in the dataset into nominal attributes. Discretization is by simple binning. Skips the class attribute if set.

Supports equal-frequency and equal-width binning.
For example, put $1, 2, 3, 4, 10, 20, 30, 41$ in 4 bins.

Equal width
$1, 2, 3, 4, 10$ in $(-\infty, 11]$
20 in $(11, 21]$,
30 in $(21, 31]$
41 in $(31, +\infty)$

Equal frequency
$1, 2$ in $(-\infty, 2.5]$
$3, 4$ in $(2.5, 7]$
$10, 20$ in $(7, 25]$
$30, 41$ in $(25, +\infty)$

MathExpression

Modify numeric attributes according to a given expression.

- The letter "A" refers to the attribute value.
- MIN, MAX, MEAN, SD refer respectively to minimum, maximum, mean and standard deviation of the attribute.
- The following operators are supported: +, -, *, /, (,), pow, log, abs, cos, exp, sqrt, tan, sin, ceil, floor, rint, MEAN, MAX, MIN, SD, COUNT, SUM, SUMSQUARED, ifelse

rint(1.3)=1, rint(1.8)=2

NominalToBinary

Converts all nominal attributes into binary numeric attributes. A nominal attribute with k values is transformed into k binary attributes (using the one-attribute-per-value approach).

Normalize

Normalizes all numeric values in the given dataset (apart from the class attribute, if set). The resulting values are by default in $[0, 1]$ for the data used to compute the normalization intervals.

Obfuscate

Renames the relation, all attribute names and all nominal (and string) attribute values. For exchanging sensitive datasets.

PrincipalComponents

Performs a principal components analysis and transformation of the data. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data—default 0.95 (95%).

Remove

Removes a range of attributes from the dataset.

See also RemoveType and RemoveUseless.

RemoveMissingValues

Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

Standardize

Standardizes all numeric attributes in the given dataset to have zero mean and unit variance (apart from the class attribute, if set).

Outline

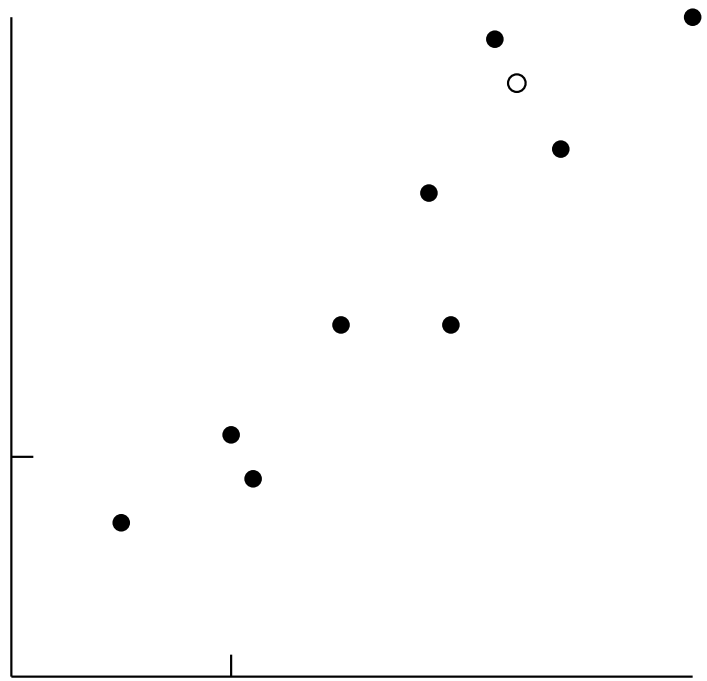
Principal Components

- Can change [orthogonal] coordinate system.
- Whatever coordinate system one uses, the sum of variances along each axis is constant.
- Maximize the variance along the first axis, then choose the second axis such that it captures as much as possible of the remaining variance, and so on.

Original data

	x	y
	2.5	2.4
	0.5	0.7
	2.2	2.9
	1.9	2.2
	3.1	3.0
○	2.3	2.7
	2	1.6
	1	1.1
	1.5	1.6
	1.1	0.9
mean	1.81	1.91
s^2	0.61 . . .	0.71 . . .

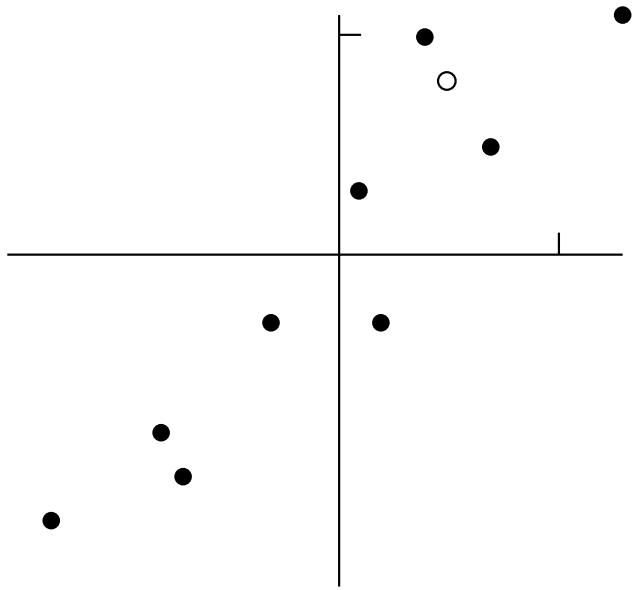
$s_x^2 + s_y^2 = 1.33\ldots$



Remove the mean

	x	y
	0.69	0.49
	−1.31	−1.21
	0.39	0.99
	0.09	0.29
	1.29	1.09
○	0.49	0.79
	0.19	−0.31
	−0.81	−0.81
	−0.31	−0.31
	−0.71	−1.01
mean	0	0
s^2	0.61 . . .	0.71 . . .

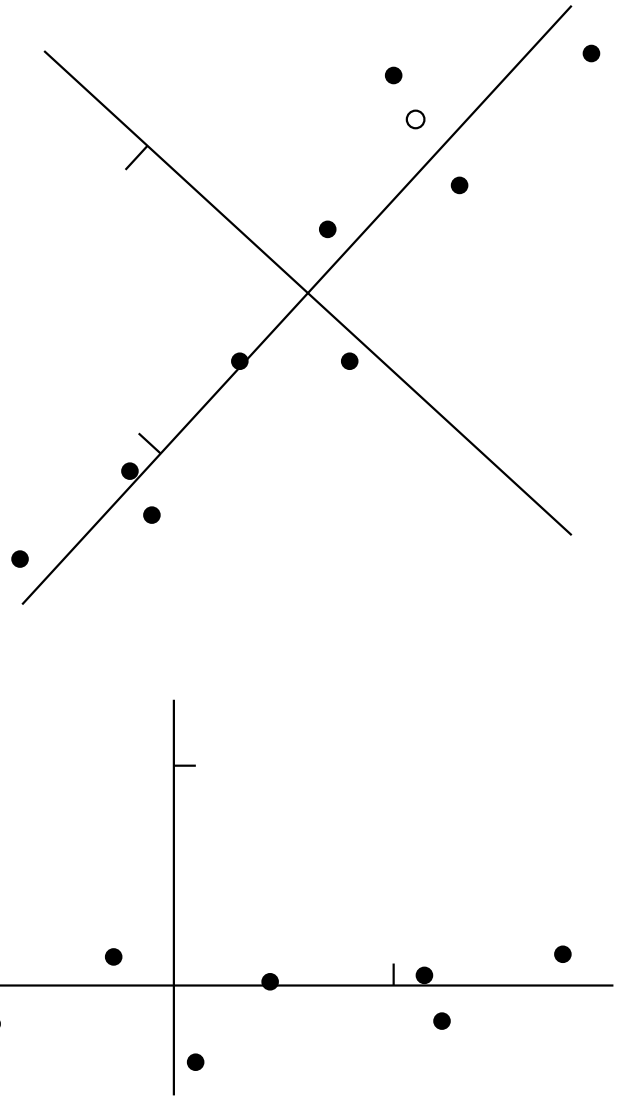
$s_x^2 + s_y^2 = 1.33\ldots$



Find principal component(s)

	x	y
	$-0.82\ldots$	$-0.17\ldots$
	$1.77\ldots$	$0.14\ldots$
	$-0.99\ldots$	$0.38\ldots$
	$-0.27\ldots$	$0.13\ldots$
	$-1.67\ldots$	$-0.20\ldots$
○	$-0.91\ldots$	$0.17\ldots$
	$0.09\ldots$	$-0.34\ldots$
	$1.14\ldots$	$0.04\ldots$
	$0.43\ldots$	$0.01\ldots$
	$1.22\ldots$	$-0.16\ldots$
mean	0	0
s^2	$1.28\ldots$	$0.04\ldots$

$s_x^2 + s_y^2 = 1.33\ldots$



Dimensionality reduction

	x
	$-0.82\ldots$
	$1.77\ldots$
	$-0.99\ldots$
	$-0.27\ldots$
	$-1.67\ldots$
○	$-0.91\ldots$
	$0.09\ldots$
	$1.14\ldots$
	$0.43\ldots$
	$1.22\ldots$
mean	0
s^2	$1.28\ldots$

Travail pratique

Changer le fichier `weather.arff` :

- la temperature doit être exprimée en degrée Celcius, en tant qu'entier ($^\circ C = (^\circ F - 32)/1.8$);
- l'humidité doit être représentée par 3 intervals;
- l'attribut `outlook` doit être représenté par des attributs binaires.