CS423: Data Mining

Data Preprocessing

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

Quote of the day

"You cannot teach a man anything; you can only help him discover it in himself."

- *Galileo* -

Outline

- Why Data Preprocessing ?
- Major tasks in data preprocessing
  - Data cleaning
  - Data integration
  - Data reduction
  - Data transformation

Why data preprocessing ?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest
    - e.g., occupation = " "
  - noisy: containing errors or outliers
    - e.g., salary = "-10"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
- Raw data is not in the right format
  - Image data, stream of data, attributes take on different range.

Why is data dirty ?

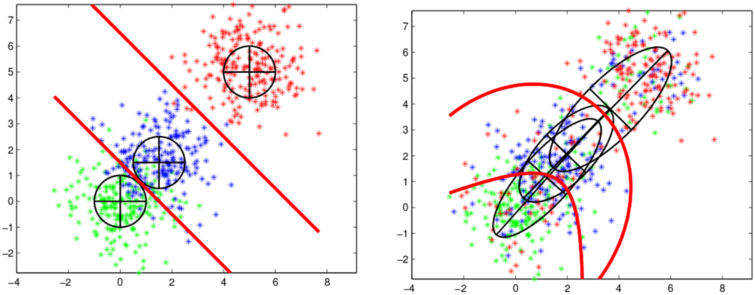
- Incomplete data may come from
  - Bad data collection protocol or plan
  - Different considerations between the time when the data was collected and when it is analyzed.
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

Why data preprocessing is important ?

- No quality data, no quality mining results
- Missing data may cause incorrect or even misleading statistics.

Example of bad data

Learning a classifier in the presence of label noise (right).



Major tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, identify or remove outliers
- Data integration
  - Integration of multiple datasets
- Data transformation
  - Normalisation and aggregation (mean, variance, etc.), feature extration.
- Data reduction
  - Obtain reduced representation in volume but produces the same analytical results.

Data cleaning

Incomplete (missing) data

- Mostly, we are interested in dealing with missing attributes
- Missing attributes may be due to
  - equipment malfunction
  - data not entered due to misunderstanding
  - certain attributes may not be considered important at the time of entry
- Missing data may need to be inferred

How to handle incomplete data ?

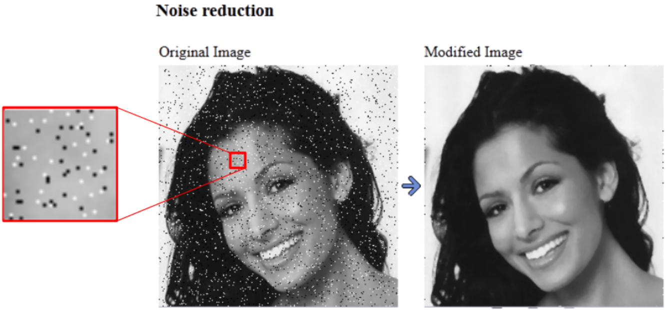
- Ignore the data point: usually done when class label is missing (when doing classification)—not effective why?
- Fill in the missing value manually: problem?
- Fill in automatically with
  - a global constant : e.g., "unknown", "0".
  - the attribute mean: good
  - the attribute mean for all samples belonging to the same class: better
  - the attribute mean for all samples within same neighbourhood: best ?

Noisy data

- Noise: random error or variance in a measured variable
- Incorrect attribute values and class label may be due to
  - Faulty data collection instruments
  - Disagreement between data experts
  - Data transmission problems
  - Technology limitation

How to handle noisy data ?: Binning

- Used to reduce the effects of minor observation error by consulting neighbours.
- then one can smooth by bin means, smooth by bin median, etc.



How to handle noisy data ?: Clustering

- Used to detect and remove outliers

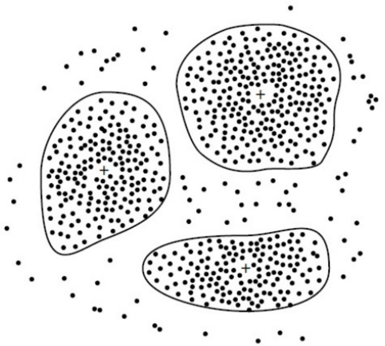
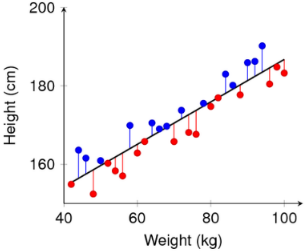


Figure 1.11 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster "center" is marked with a "+".

How to handle noisy data ?: Regression

- Used to detect and remove outliers or smooth out the data.
- Require knowledge of relationships between attributes.



How to handle noisy data ?: Consulting human expert

- Used to detect suspicious values.
- Used when working on sensitive data.
- Often more expensive to perform in terms of time, labour, and money.

Data integration

Data integration

- Combines data from multiple sources into a coherent store.
- Data integration introduces
  - Attribute redundancy (for example merging two tables with slightly different attribute names but with same value)

Handling redundancy

- Redundant data occur often when integration of multiple databases
- Redundant attributes may be able to be detected by correlation analysis (for nominal data) and covariance analysis (for numeric data)
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

- We want to find out if $x^1 \in \{a, b, \dots\}$ and $x^2 \in \{\alpha, \beta, \dots\}$ are independent.
- If they are independent, we keep both. Otherwise we drop one of them.
- We can use χ^2 test, which is a form of hypothesis testing
- Our null hypothesis is that two attributes/features are independent

Correlation Analysis (Example data)

Results from a survey asking if students love to play chess and read science fiction.

	Play chess	Not play chess	Sum (row)
Like science fiction	250	200	450
Not like science fiction	50	1000	1050
Sum (col.)	300	1200	1500

$x^1 \in \{\text{'chess'}, \text{'not chess'}\}$ and $x^2 \in \{\text{'science fiction'}, \text{'not science fiction'}\}$

Are x^1 and x^2 correlate ?

Correlation Analysis (Nominal Data)

- Chi-square computes the deviation of the observed frequencies of events from the expected frequencies of events

$$\chi^2 = \sum_{i=0}^{C} \sum_{j=0}^{R} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

- The larger the χ^2 value the more sure we are to reject null hypothesis.
- According to the null hypothesis, the expected frequency is based on independence assumption, $P(a, \beta) = P(a) \times P(\beta)$
- So the expected frequency is calculated as

$$E_{a,\beta} = P(a) \times P(\beta) \times sampleSize$$

Correlation Analysis (Example data)

Results from a survey asking if students love to play chess and read science fiction.

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200	450
Not like science fiction	50	1000	1050
Sum (col.)	300	1200	1500

Computing $E_{chess,fiction}$ we find

$$\begin{aligned} E_{chess,fiction} &= P(chess) \times P(fiction) \times samplesize \\ &= \frac{300}{1500} \times \frac{450}{1500} \times 1500 = 90 \end{aligned}$$

Correlation Analysis (Example data)

Results from a survey asking if students love to play chess and read science fiction.

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum (col.)	300	1200	1500

Computing χ^2 we find

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} \approx 507.93$$

It shows that playing chess and liking science fiction are correlated

Covariance (Numeric data)

$$cov(x^k, x^m) = \frac{\sum_{i=1}^{N}(x^k - \mu^k)(x^m - \mu^m)}{N}$$

- N is number of data points, μ^k, μ^m are the mean of feature k and m.
- If $k = m$ then the formula reduces to variance.
- Covariance measures how much two features change together.
- If features are independent we will find $cov(x^k, x^m) = 0$
  - But when covariance = 0 but are not independent.
  - Covariance measure *linear* dependency.

Covariance (Example)

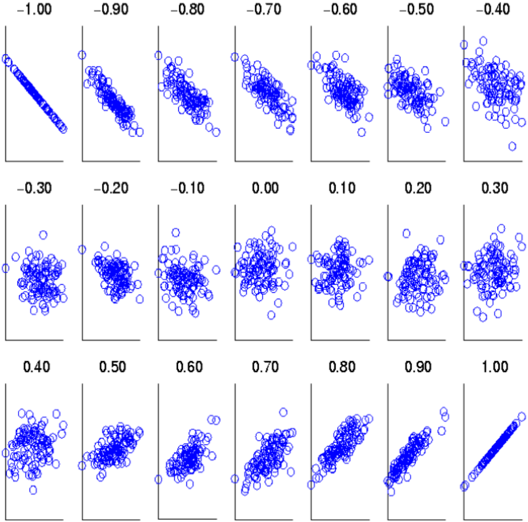
- Suppose two stocks x and y have the following values in one week.
  - (2,5), (3,8), (5, 10), (4, 11), (6,14)
- if the stocks are affected by the same industry trends, will their prices rise or fall together
  - $mu_x = 4,\ mu_k = 9.6$
- $cov(x, y) = ((2-4)(5-9.6) + (3-4)(8-9.6) + (5-4)(10-9.6) + (6-4)(14-9.6))/5 = 5.92$
- Thus x, y rise together.

Correlation Analysis

Correlation coefficient a.k.a. Pearson's product moment coefficient is a normalised covariance

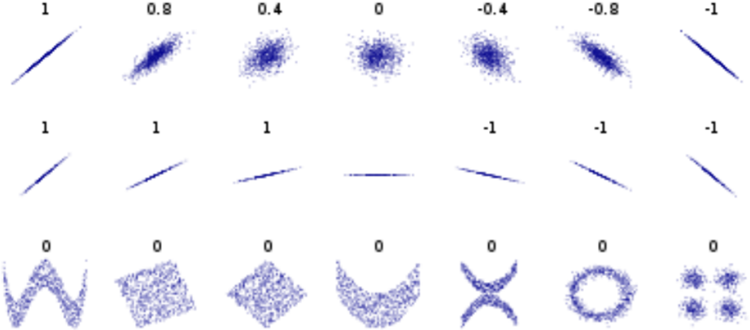
- $R(x, y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$
- Measures linear correlation
- $R(x, y) > 0$, x and y are positively correlated.
- $R(x, y) = 0$: linearly independent.
- $R(x, y) < 0$: negatively correlated.

Visualising Correlation



Scatter plots showing the correlation from –1 to 1.

Visualising Correlations



Data reduction

Definition

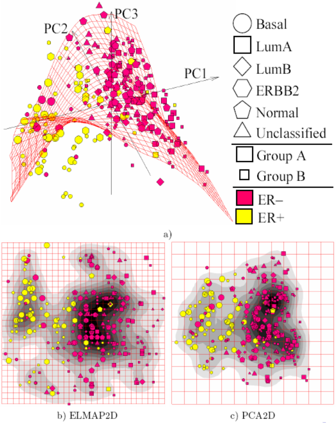
- To obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why ? A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run.
- Two strategies
  - Dataset size reduction
    - ⋆ Histogram, Clustering, Sampling, Model-based (Mixture model)
  - Dimensionality reduction
    - ⋆ Principal component analysis, Feature subset selection.

Dimensionality Reduction

- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful.
  - The possible decision hyperplane will grow exponentially.
- Dimensionality reduction
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualisation
- Dimensionality reduction techniques
  - Principal Component Analysis

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.



Feature (Subset) Selection

- Select relevant subset of features for the task.
- Can be divided into 3 categories.
  - Wrapper: Find all combinations of features and evaluate the usefulness of the subset using task's performance criteria. (e.g., the predictive performance)
  - Filter: Most general, use some criteria to filter out redundant features.
  - Embedded
    - ★ Feature selection is part of the algorithm.
    - ★ Regularisation

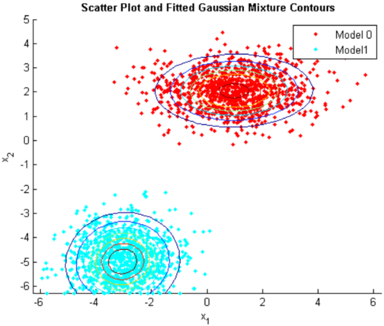
Data reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data
  - E.g., Mixture model
- Non-parametric methods
  - Do not assume data models.
  - Major families: histograms, clustering, sampling

Gaussian Mixture Model

- Assume the data obeys Gaussian Distribution, Estimate the model and store only the model's parameters.

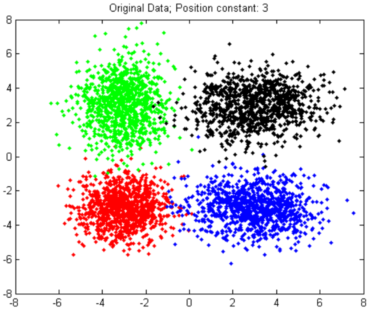
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only

- Can be very effective if data is clustered but not if data is “smeared”



Sampling

- Sampling: obtaining a small sample s to represent the whole data set N
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling.

Types of Sampling

- Simple random sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - Once an object is selected, it is removed from the population
- Sampling with replacement
  - A selected object is not removed from the population
- Stratified sampling:
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Data transformation

Data transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values.
- Tasks
  - Attribute/feature construction
    - New attributes constructed from the given ones.
    - Image -> data point, Signal -> data point
  - Normalization: Scaled to fall within a smaller, specified range
    - Min-max normalisation
    - z-score normalisation
    - normalisation by decimal scaling
  - Discretisation: : Transform continuous variable to discrete ones.

Image to data point

Reshaping an image into a vector of length (Width x Height)

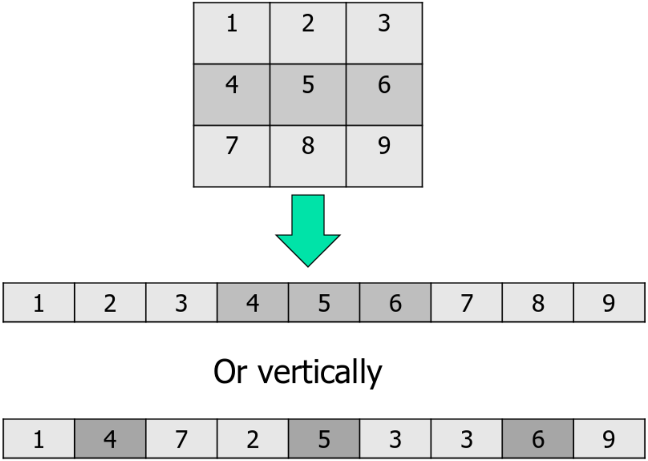


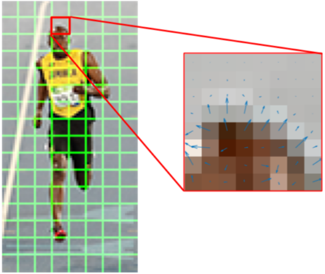
Image: histogram of gradients

Useful for extracting edges, corners from images. Steps

1. Calculating gradients in x and y direction by central difference

$$g_x = \frac{I(i-1,j) - I(i+1,j)}{2} \text{ and } g_y = \frac{I(i,j-1) - I(i,j+1)}{2} \text{ (might need padding)}$$

2. Final gradient is $g = \sqrt{g_x^2 + g_y^2}$ and its direction is $\theta = \arctan \frac{g_y}{g_x}$
   - We will get two additional matrix of gradient magnitude and direction



2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

Gradient Magnitude

80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

Gradient Direction

Image: histogram of gradients

3. Construct a histogram of based on directions.

4. Sum of magnitudes which point to the same direction contributes to the frequency of that direction.

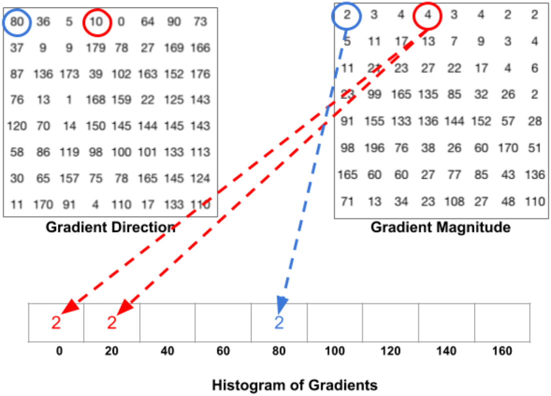


Image texture extraction: Local Binary Pattern

LBP is used to extract texture information from images. The steps are

1. Convert image to grayscale

2. Traverse along all pixels

3. For each pixel, consider its neighbourhood of size r

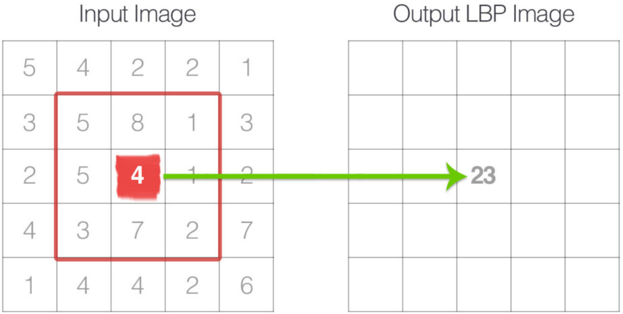


Image texture extraction: Local Binary Pattern

4. We threshold neighouring pixels and write 0 if the neighbour is greater than the centre pixel and 1 otherwise.
   - For $r = 1$ there will be 8 neighbour so the above thresholding will produce 8-bits string.

5. Evaluate the bitstring gives LBP value of the centre pixel.

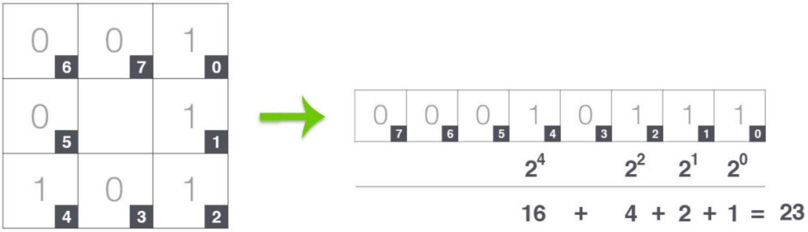


Image texture extraction: Local Binary Pattern

6. Calculate histogram of values in LBP image
   - For $r = 1$, we get 8-bits string with value from 0 to 255, so we will construct a histogram of 256-bin.

7. The resulting histogram is the texture feature of the image.

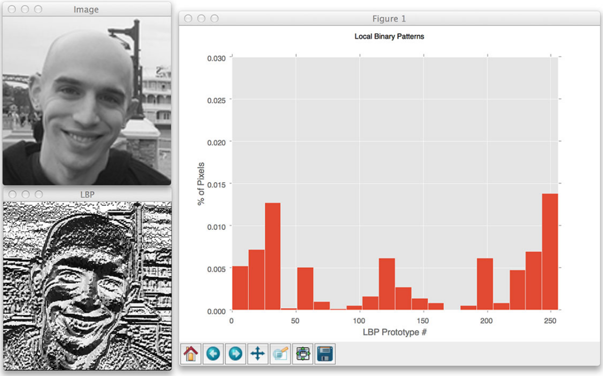
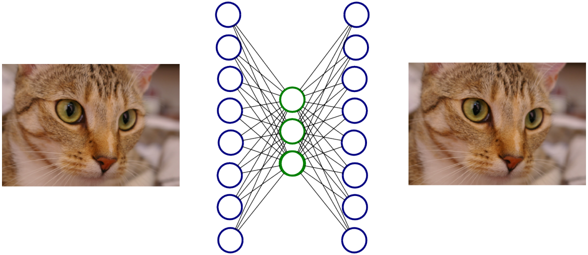


Image feature extraction: Autoencoder

- An artificial neural network used for unsupervised learning of efficient codings.
- The aim of an autoencoder is to learn a representation (encoding) for a set of data.



Normalisation

- Min-max normalisation

$$x' = \frac{x - max_x}{max_x - min_x}(newMax_x - newMin_x) + newMin_x$$

- Z-score normalisation

$$x' = \frac{x - \mu_x}{\sigma_x}$$

- Decimal scaling

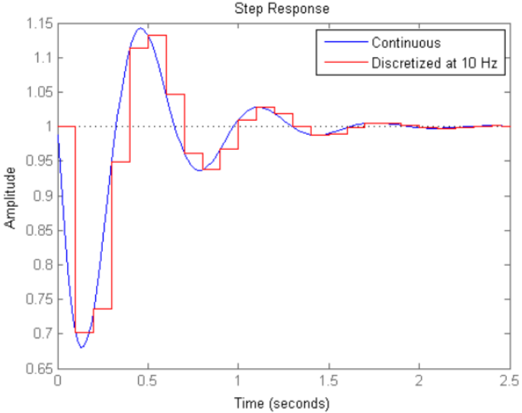
$$x' = \frac{x}{10^j} \quad \text{where } j \text{ is the smallest integer s.t. } \max(|x'|) \leq 1$$

Discretisation

- Discretisation: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Discretization can be performed recursively on an attribute
- Typical methods
  - Binning
  - Histogram analysis
  - Clustering analysis
  - Using knowledge from expert

Sound -> Data vector

Discretising stream of data



References

- Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.
- Ojala, Timo, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on pattern analysis and machine intelligence 24.7 (2002): 971-987.
- Ramírez-Gallego, Sergio, et al. "Data discretization: taxonomy and big data challenge." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 6.1 (2016): 5-21. `http://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1996_2015-sramirez-taxonomy.pdf`