Some Clustering Methods on Dissimilarity or Similarity Matrices:
Uncovering Clusters in WEB Content, Structure and Usage

Yves Lechevallier
INRIA-Paris-Rocquencourt
AxIS Project

Yves.Lechevallier@inria.fr

INRIA
Paris-Rocquencourt

Workshop Franco-Brasileiro sobre Mineração de Dados
Workshop Franco-Brésilien sur la fouille de données
Récife 5-7 May 2009

Two types of Data Tables

Variable	Moyenne	Ecart-type	Minimum	Maximum
CA	102.46	118.92	1.20	528.00
MG	25.86	28.05	0.20	95.00
NA	93.85	195.51	0.80	968.00
K	11.09	24.22	0.00	130.00
SUL	135.66	326.31	1.10	1371.00
NO3	3.83	6.61	0.00	35.60
HCO3	442.17	602.94	4.90	3380.51
CL	52.47	141.99	0.30	982.00

Tableau 1.3. *Statistiques sommaires des variables continues*

Classical Data Table

Each object is described by a vector of measures.

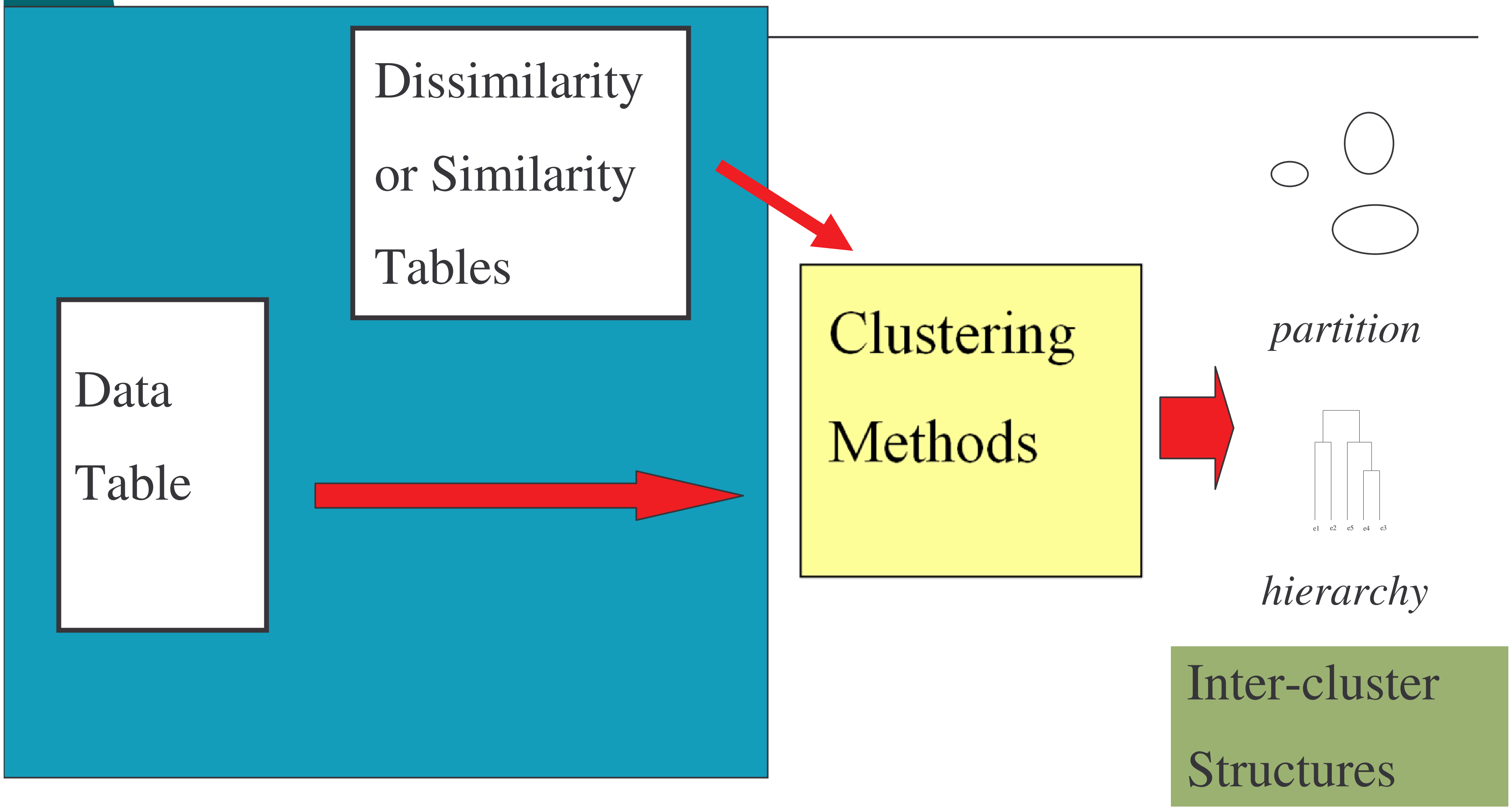
	CA	MG	NA	K	SUL	NO3	HCO3	CL
CA	1.00							
MG	0.70	1.00						
NA	0.12	0.61	1.00					
K	0.13	0.66	0.84	1.00				
SUL	0.91	0.61	0.06	-0.03	1.00			
NO3	-0.06	-0.21	-0.12	-0.17	-0.16	1.00		
HCO3	0.13	0.62	0.86	0.88	-0.07	-0.06	1.00	
CL	0.28	0.48	0.59	0.40	0.32	-0.12	0.19	1.00

Tableau 1.4. *Matrice des corrélations*

Dissimilarity or Similarity Table

The relation between two objects is measured by a positive value.

Clustering Process



Components of a Clustering Problem

To formulate a clustering problem you must specify the following components

- Ω : the set of objects (units) to be clustered.
- The set of variables (attributes) to be used in describing objects.
- A principle for grouping objects into clusters (based on a measure of similarity or dissimilarity between two objects)
- The inter-cluster structure which defines the desired relationship among clusters (clusters should be disjoint or hierarchically organised)

Partitioning Methods

The selected inter-cluster structure is the partition.

By defining a function of homogeneity or a quality criterion on a partition, the problem of clustering becomes a problem perfectly defined in discrete optimization.

To find, among the set of all possible partitions, a partition where a fixed a priori criterion is optimized.

Optimisation problem

A criterion W on $\wp_K(\Omega) \rightarrow \Re^+$, where $\wp_K(\Omega)$ is a set of all partitions in K nonempty classes of Ω that the problem of optimization is :

$$W(P) = \underset{Q \in \wp_K(\Omega)}{Min} W(Q) = \sum_{k=1}^{K} w(Q_k)$$

where $w(Q_k)$ is the homogeneity measure of the class Q_k.

and K is the number of classes

Iterative Optimization Algorithm

- We start from a realizable solution $Q^{(0)} \in \wp_K(\Omega)$

  Choice

- At the step *t+1*, we have a realizable solution $Q^{(t)}$

  we seek a realizable solution $Q^{(t+1)} = g(Q^{(t)})$

  checking $W(Q^{(t+1)}) \leq W(Q^{(t)})$

  Choice

- The algorithm is stopped when $Q^{(t+1)} = Q^{(t)}$

Remark : the probability to obtain one best solution is $1-(1-p)^B$ where B is the number of runs and *p* is the probability to obtain one best solution for each initial solution.

Neighborhood algorithm

One of the strategies used to build the function *g* is :

- to associate any realizable solution *Q* a finite set of the realizable solutions *V(Q),* call *neighborhood* of *Q*,
- then to select the optimal solution for this criterion *W* in this neighbour, which is usually called local optimal solution.

For example we can take as neighborhood of *Q* all partitions obtained starting from the partition *Q* by changing only one element of class.

Two well known exemples of this algorithm are « ping pong » algorithm and *k-means* algorithm.

k-means algorithm

With the neighborhood algorithm, it is not necessary systematically to take a best solution to obtain the decrease of the criterion,

it is sufficient to find in this neighborhood a solution better than the current solution. In the *k-means* algorithm it is sufficient:

to determine ℓ such as $\ell = \arg\min_{j=1,\cdots,K} d^2(\mathbf{z}_i, \mathbf{w}_j)$

The decrease of the intraclass inertia criterion W is ensure thanks to the Huygens theorem.

Iterative two steps relocation process

This algorithm involves at each iteration two steps:

1. The first step is the representation step. The goal is to select a prototype for each cluster by optimizing an a priori criterion.
2. The second step is the allocation step. The goal is to find a new affection of each object of Ω from prototypes defined in the previous step.

Dynamic Clustering Method

Dynamical clustering algorithms are iterative two steps relocation algorithms involving at each iteration the identification of a prototype for each cluster by optimizing an adequacy criterion.

It is a *k-means* like algorithm with adequacy criterion equal to variance criterion and the class prototypes equal to cluster centers of gravity

Optimization problem

In dynamical clustering, the optimization problem is :

Let Ω be a set of *n* objects described by *p* variables and Λ a set of class prototypes.

Each object *i* is described by a vector $\mathbf{x}_i$.

The problem is to find simultaneously the partition $P=(C_1,...,C_K)$ of Ω in K clusters and the system $L=(L_1,...,L_K)$ of class prototypes of Λ which optimize the partitioning criterion W(*P*,*L*).

$$W(P,L) = \sum_{k=1}^{K} \sum_{s \in C_i} D(\mathbf{x}_s, L_k) \quad C_k \in P, L_k \in \Lambda$$

Algorithm

(a) Initialization

Choose K distinct class prototypes $L_1,\ldots,L_K$ of Λ

(b) Allocation step

For each object i of Ω define the index cluster l which verifies

$$l = \arg\min_{k=1,\ldots,K} D(\mathbf{x}_i, L_k)$$

(c) Representation step

For each cluster k find the class prototype L_k of Λ which minimizes

$$w(C_k, L) = \sum_{s \in C_k} D(\mathbf{x}_s, L)$$

Repeat (b) and (c) until the stationarity of the criterion

Convergence

In order to get the convergence it is necessary to define the class prototype L_k which minimizes the adequacy criterion $w(C_k, L_k)$ measuring the proximity between the prototype L_k and the corresponding cluster C_k

↓

- The dynamical clustering algorithm converges
- The partitioning criterion decreases at each iteration

How to define D ?

The optimization problem for class prototype

For each cluster *C* we search the vector *L* of *S* which minimizes the following adequacy criterion :

$$w(C,L) = \sum_{s \in C} D(\mathbf{x}_s, L) = \sum_{s \in C} d^2(\mathbf{x}_s, L) = \sum_{j=1}^{p} \sum_{s \in C} (x_s^j - L^j)^2 \quad L^j \in \Re$$

For each variable *j* the problem is to find the element L^j of *S* which minimizes:

$$\sum_{s \in C} (x_s^j - L^j)^2$$

The solution is evident $L^j = \frac{1}{|C|} \sum_{s \in C} x_s^j$

Two Classical Criteria

d is euclidian distance

Λ is $\Re$.

$D = d^2$

$$W(P,L) = \sum_{k=1}^{K} \sum_{s \in C_i} d^2(\mathbf{x}_s, L_k)$$

Mean vector

Unique

$$L_k^j = \frac{1}{|C|} \sum_{s \in C} x_s^j$$

$D = d$

$$W(P,L) = \sum_{k=1}^{K} \sum_{s \in C_i} d(\mathbf{x}_s, L_k)$$

Median vector

No unique

$$L_k^j = median\{x_s^j, s \in C_k\}$$

How to classify the Complex Data

Three major approaches:

- **Vectorial translation** (vectorial model of Salton for the analysis of texts)
  - losses of information, distortion of coding,…
- **Construction of tables of proximities**
  - **Problem** : Choose of the measure
  - **Avantage** : Use of generic tools
- **Construction of specific tools specific on the type of complex data**
  - Interval vectors
  - Funtional data,…

The optimization problem for the distance table

For each cluster C we search the object s_C of E which minimizes the following adequacy criterion :

$$w(C_k, s') = \sum_{s \in C_k} d^2(s, s')$$

The solution is simple

$$s_{C_k} = \arg\min_{s' \in E} \sum_{s \in C_k} d^2(s, s')$$

Clickstream data

- Web site analyzed: Informatics' Center (CIn) from Recife/Brazil
- This web site is based on dynamic pages implemented by Java servlets
- The site is quite small and well organized :
  - 91 pages
  - the site map is a tree of depth 5
- The web log ranges from June 26th 2002 to June 26th 2003 :
  - it corresponds to about 2Go of raw data
  - after the pre-processing and cleaning steps, these data represent 113 784 navigations.

The welcome page of the site

Motivation

- The classification of pages from a Web site can put in evidence the suitability between:
  - the real structure of the site (semantic structure)
  - the practice of the users and its use (the navigation or visits of the users)

- The application of a dissimilarity measure allows the classification of pages
  - the main problem: what measure among many possibilities, is the most appropriated?

Semantic structure of the site

Great density of links

- **Home**
- **The Informatics' center**
  - Mission
  - Objectives
    - In the informatics' center
    - In the university
    - In the region
    - In the country
    - In the world
  - Organization chart
  - Departments
  - Courses
  - Resolutions
  - Localization
  - Tunnel of time
- **People**
  - Direction
  - Coordinators
  - Professors
  - Support
  - Management
  - Administrative
- **Graduation**
  - Information of the graduation
  - How to be a student
  - How to ingress in the pos-graduation
  - Coordination/Secretariat
- **Post-graduation**
  - Registrations and Information
    - Subjects for Master 2004
    - Address
    - Documents
      - New internal regiment
      - Internal regiment (11/1996)
      - Internal regiment (02/1992)
      - Qualification norms and these proposition
      - School registration 2/2000
      - School registration 1/2000
      - School registration 1/2001
    - Summary of disciplines
    - Masters results
    - School registration and pre-registration
    - Orientation of School registration
    - PhD results
  - Description
  - Professors group
  - Lines of research
  - Specialization
    - Information
      - Concentration areas
      - Course's Structure
    - Professors group
    - Computational Resources
  - Masters
  - PhD
  - APG
  - Administrative
  - Financial resources
  - Events
- **Extension**
  - Partners
  - Events
  - Services
  - UFPE for all
- **Research**
  - Groups of research
  - Publications
    - Theses and dissertations
    - Technician reports
    - Tutorials, manuals and class notes
    - Book chapters
    - Complete books
    - Complete papers in periodic
    - Complete Work in proceedings
    - Summary in proceedings
  - Projects of research
  - Events
    - Events related to the Informatics' center
    - Events of interest
  - Honors and prizes
- **Infrastructure**
- **Services**
- **News**
  - Releases
  - Clipping
  - Contacts
- **Phones**
  - Professors
  - Sectors
  - Employees

Navigation or Visite

A navigation is a set of clicks performed by an user during a period of time.

The end of this period corresponds to the absence of clicks during at least 30 minutes.

One request is associated to each click and corresponds to one requisition for a page in the web site.

Two complex representations of navigations

Site :

{A,B,C,D} 4 pages

two navigations

n_1=(A,B,A,C,D)
n_2=(A,B,C,B)

	n_1	n_2
A	{1,3}	{1}
B	{2}	{2,4}
C	{4}	{3}
D	{5}	

Choice of the dissimilarity function

Jaccard

binary

$$ {}_J(p_i, p_j) = \frac{|\{k|n_{ik} \neq n_{jk}\}|}{|\{k|n_{ik} \neq 0 \text{ ou } n_{jk} \neq 0} $$

Cosine

counting

$$ d_{\cos}(p_i, p_j) = 1 - \frac{\sum_{k=1}^{N} m_{ik} m_{jk}}{\sqrt{\left(\sum_{k=1}^{N} m_{ik}^2\right)\left(\sum_{k=1}^{N} m_{jk}^2\right)}} $$

Tf x idf

counting

$$ d_{\mathrm{tf} \times \mathrm{idf}}(p_i, p_j) = 1 - \sum_{k=1}^{N} w_{ik} w_{jk}, $$

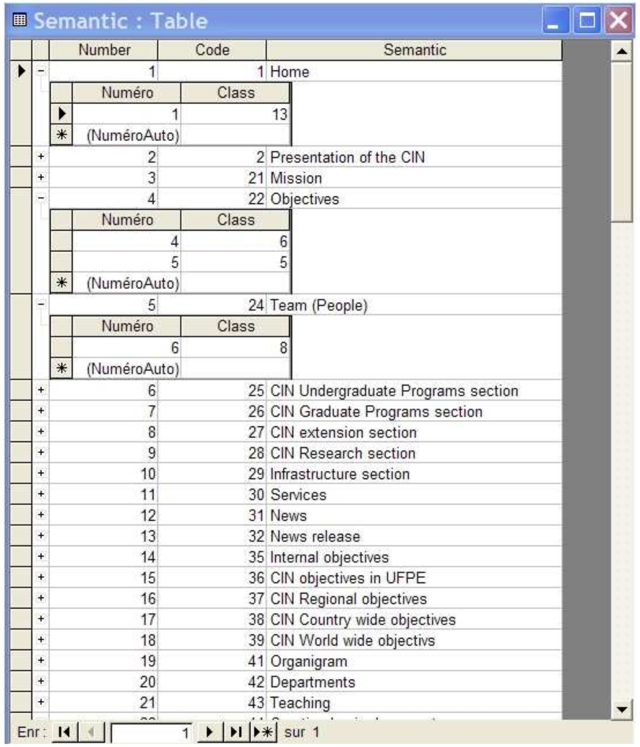
with

$$ \text{avec } w_{ik} = \frac{m_{ik} \log \frac{P}{P_k}}{\sqrt{\sum_{l=1}^{N} m_{il}^2 \log \left(\frac{P}{P_l}\right)^2}} $$

These three measures don't integrate the semantic structure in the computation

Expert or a priori partition

1	2	3	4	5	6	7
Publications	Research	Partners	Undergraduate	Objectives	Presentation	Directory
8	9	10	11	12	13	
Team	Options	Archives	Graduate	News	Others	



Classification of pages into semantic categories performed by an expert.

Results on the distance table

The dynamic clustering :

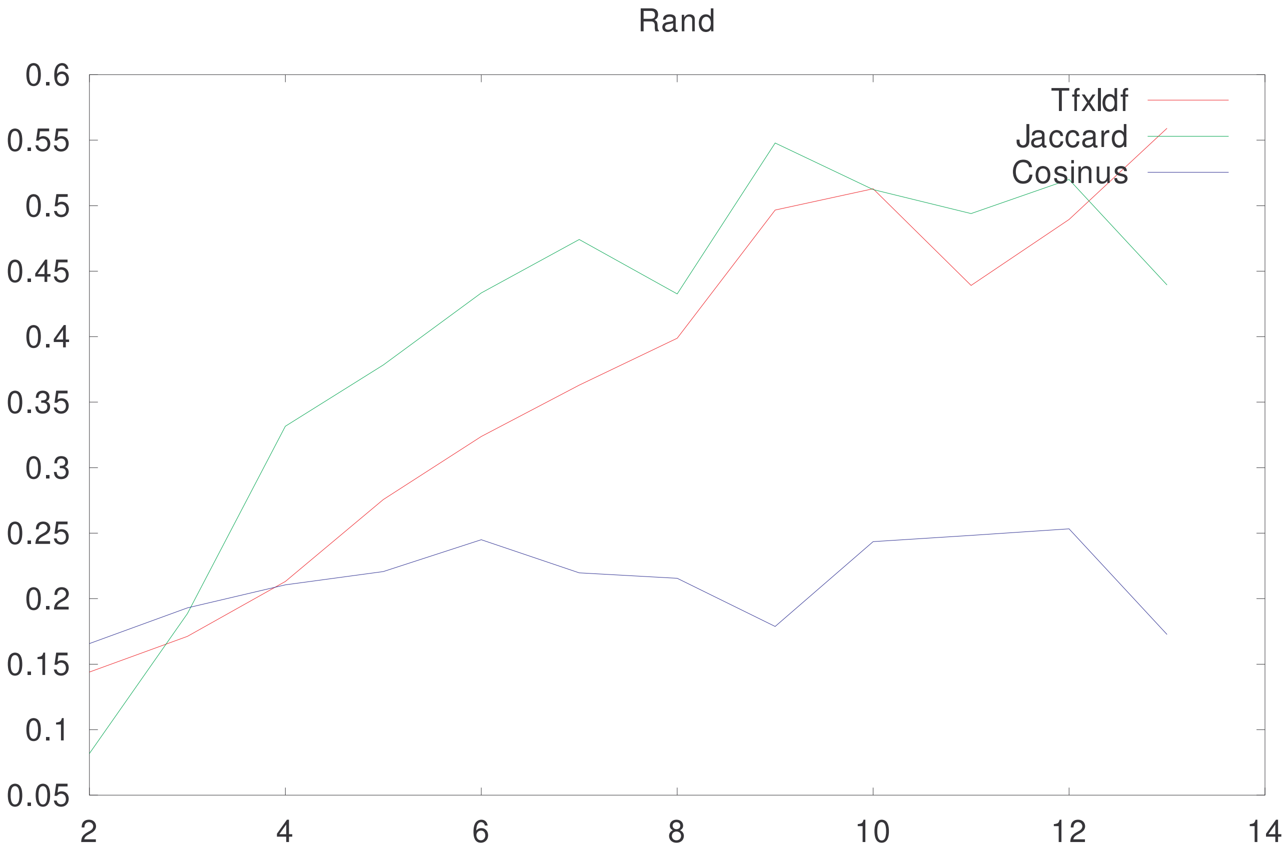
Dissimilarity	Rand index	Found classes	min F mesure
Jaccard	0.5698 (9 classes)	6	0.4444
Tf×idf	0.5789 (16 classes)	7	0.5
Cosinus	0.3422 (16 classes)	4	0.3

Hierarchical clustering :

Dissimilarity	Rand index	Found classes	min F mesure
Jaccard	0.6757 (11 classes)	3	0.5
Tf×idf	0.4441 (15 classes)	3	0.4
Cosinus	0.2659 (11 classes)	5	0.4

External evaluation by Rand and F measure criteria

Rand / MND on distance table



A dynamical cluster method with adaptive distances (G. Govaert, 1975)

$d = (d_1, \ldots, d_K)$ is a vector of K distances. The distance d_k, is associated to the cluster C_k and belongs to the set of distance family D.

Our method searches a pair (P, L) and a vector d of D^K which optimize the criterion $W_2(P,L,d)$.

$$W_2(P,L,d) = \sum_{k=1}^{K} \Delta_2(C_k, y_k, d_k) = \sum_{k=1}^{K} \sum_{i \in C_k} d_k^2(\mathrm{x}_i, \mathrm{y}_k)$$

d_k is a distance of the cluster C_k (local allocation distance)

y_k is the prototype of the cluster C_k

The optimization problem for the distance table

$L=(c_1,..,c_K)$ is a vector of objects of Ω

$\lambda = (\lambda_1, \ldots, \lambda_K)$ a weight vector λ of the partition P

with the contraints $\prod_{k=1}^{K} \lambda_k = 1 \text{ and } \lambda_k > 0 \text{ for } k = 1, \ldots, K$

Our method searches a pair (P, L) and a weight vector λ where the criterion $W_2(P,L,\lambda)$ is optimized

$$W_2(P, L, \lambda) = \sum_{k=1}^{K} \Delta_2(C_k, c_k, \lambda_k) = \sum_{k=1}^{K} \sum_{s \in C_k} \lambda_k d^2(s, c_k)$$

The optimization problem of the representative step

The optimization problem of the representative step
Is divided in two steps:

Step 1: The class C_k and weight λ_k are fixed.

For each cluster C_k, the problem is to find a object c_k that minimizes the adequacy criterion

$$c_k = \arg\min_{c \in \Omega} \Delta_2(C_k, c, \lambda_k) = \arg\min_{c \in \Omega} \sum_{s \in C_k} \lambda_k d^2(s, c)$$

The optimization problem of the representative step

Step 2: The partition P and $L=(c_1,..,c_K)$ is a vector of objects are fixed.

The problem is to find weight vector λ that minimizes the adequacy criterion

$$W_2(P,L,\lambda) = \sum_{k=1}^{K} \sum_{s \in C_k} \lambda_k d^2(s,c_k) = \sum_{k=1}^{K} \lambda_k \sum_{s \in C_k} d^2(s,c_k) = \sum_{k=1}^{K} \lambda_k \Phi_k$$

The solution is given by the Lagrange multiplier method.

Conclusion

- The adaption of the class prototype approach to classify a distance table is easy.
- The protype is remplaced by a medoid.
- This approach can be used when the distance is non an euclidean distance.

References

- **M. Chavent, F. A. T. De Carvalho, Y. Lechevallier and R. Verde**. *New clustering methods for interval data*. In Computational Statistics, Vol. 21(23):211-230, 2006.
- **A. Da Silva, Y. Lechevallier, F. Rossi and F. A. T. De Carvalho** *Clustering Dynamic Web Usage Data*. In "Innovative Applications in Data Mining",Edited by Nadia Nedjah, Luiza de Macedo Mourelle and Janusz Kacprzyk. Springer, 2009.
- **F. A. T. de Carvalho and Y. Lechevallier** *Partitional Clustering Algorithms for Symbolic Interval Data based on Single Adaptive Distances*. Pattern Recognition, 2009
- **T. Despeyroux, Y. Lechevallier, B. Trousse and A.-M. Vercoustre.** *Experiments in Clustering Homogeneous XML Documents to Validate an Existing Typology*. Journal of Universal Computer Science,2006.
- **F. Rossi, F. A. T. De Carvalho, Y. Lechevallier and A. Da Silva**. *Dissimilarities for Web Usage Mining*. In V.Batagelj, H-H. Bock, A.Ferligoj and A. vZiberna editors, Data Science and Classification (Proceedings of IFCS 2006), Pages 39-46, Springer,
- **N. Villa and F. Rossi**. *A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph*. Workshop on Self-Organizing Maps (WSOM 07).