Clustering quality evaluation: a task that has to deal with a naturally noisy context

Jean-Charles LAMIREL

Synalp-Team-LORIA
University of Strasbourg

LabelNoise-2017

Cluster quality evaluation
Challenge and principles

- In modern data analysis one central problem is related to the increasing size of data to be exploited,
- Ground truth become unavailable in most cases and the number of “categories” inherent to data must be highlighted through the use of clustering methods,
- Detection of optimal clustering model relies itself on the exploitation of clustering quality evaluation and thus on quality indexes,
- Most of the exploited techniques are based on adaptation of mean square error optimization and Euclidean distance,
- Reliability of such indexes remains an open challenge,
- Clustering is explicitly a noisy context as compared to classification.

Cluster quality evaluation

Usual distance-based indexes

- Dunn index (DU) [Dunn 74]

$$DU_k = min_{i=1,\dots,k}\left\{ min_{j=i+1,\dots,k}\left\{ \frac{diss(c_i, c_j)}{max_{m=1,\dots,k, i\neq j}\, diam(c_m)} \right\} \right\}$$

Dunn index is a diameter-based index that put the prior on models with compact and well-separated clusters. Computation time is high.

- Davis-Bouldin index (DB) [Davis 79]

$$DB_k = \frac{1}{k}\sum_{i=1}^{k} max_{j=1,\dots,k, i\neq j}\left\{ \frac{diam(c_i) + diam(c_j)}{\|c_i - c_j\|} \right\}$$

Similarly to DU index, DB index highlight models with compact and well separated cluster. It does not focuses on boundaries and is easier to compute than DU index.

Cluster quality evaluation

Usual distance-based indexes

- Calinski-Harabasz index (CH) [Calinski 74]:

$$CH_k = \frac{(N-k)}{(k-1)} \frac{BGSS}{WGSS}$$

CH is identical to variance ratio exploited in ANOVA.

- Xie-Beni index (XI) [Xie 91] is a compromise between CH and DU index. It is often used for fuzzy clustering.
- Silhouette index (SI) [Rouseeuw 87]:

$$SI_k = \frac{1}{k} \sum_{k=1}^{n} \frac{b_i - a_i}{\max(a_i, b_i)}$$

SI index takes inspiration from approaches based on modularity [Newman 06] and on nearest-neighbors. A negative value of SI means a majority of data are affected to the wrong cluster.

Cluster quality evaluation
Other indexes (1)

- Other index alternatives are based on entropy, like [Lago-Fernãndez 09] exploiting negentropy: gap between cluster entropy and the entropy of the normal distribution with the same covariance matrix,
- Graph-based measures [Pal and Biwas 97] exploit graphs of relationships between data, like relative neighborhood graphs, Gabriel graphs or spanning trees, and generalize the Dunn and Davis-Bouldin indexes to graphs to evaluate clustering quality,
- AIC [Akaike 74] and BIC [Schwarz 78] penalize the model complexity and are based on likelihood. They are expressed as :

$$AIC = argmin_k(2.\ln(L(k)) + 2.q(k))$$

$$BIC = argmin_k(2.\ln(L(k)) + q(k).\ln(n))$$

Likelihood can be estimated using WGSS [Manning et al. 08] and *q(k)* can be set to *2pk (p being the number of dim. of data).*

Cluster quality evaluation
Other indexes

- Subsampling [Ben-Hur et al. 09] consist in observing the decrease of correlation of pairs of data belonging to same clusters after generation of clustering models of same size on different data subsamples,
- Most experiments based on these alternatives are made on low dimensional data or approach needs complex parameter settings, or even complex computation, as mentioned in [Yanchi 10].

Cluster quality evaluation

Pending problems

- Behavior of indexes is analyzed on low dimensional problems and results are often contradictory
  [Liu 2011],
- Min-square error optimization have been proven to be unable to solve complex clustering problems
  [Lamirel 2011],
- Min-square and Euclidean distance based indexes are unable to produce optimal results in high dimensional context (CH and DB)
  [Kassab 2006] [Ghribi 2010],
- Most of the realistic problems are not low dimensional problems with well-shaped clusters with more or less low overlap,
- Clustering methods are imperfect and error-prone,
- Indexes results depends on the clustering methods and are not user-oriented
  [Lamirel 2004].

An alternative approach: the feature maximization metric

[Lamirel 08]

Let us consider a partition C resulting from a grouping method applied on a set of data D represented with a set of descriptive features F, feature maximization is a metric which favors groups with maximum *Feature F-measure* which represents the harmonic mean between :

Feature Recall

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_d^f} \equiv P(c|f)$$

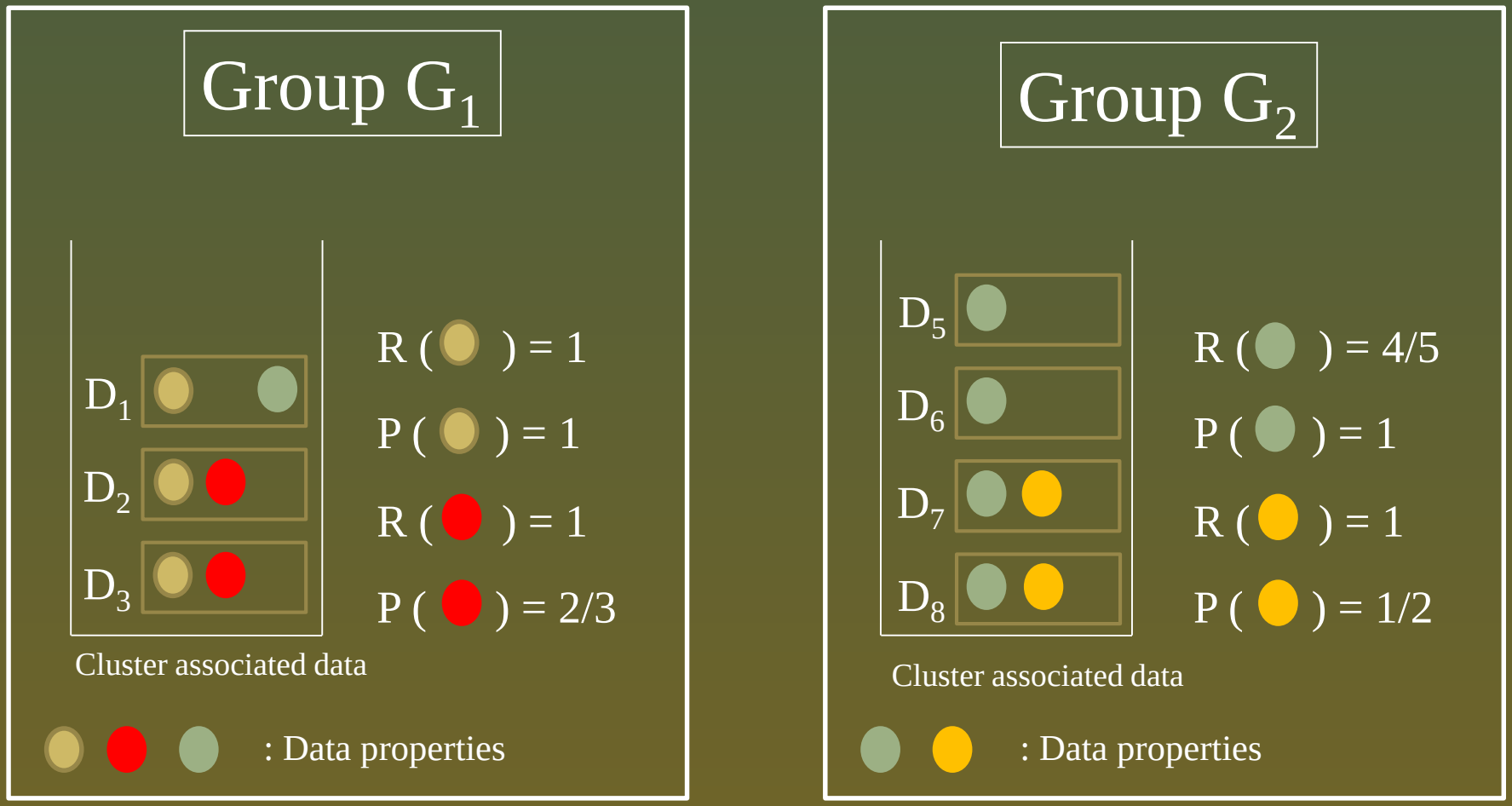
Feature Dominance

$$FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F^c, d \in c} W_d^{f'}} \equiv P(f|c)$$

A **maximized group feature** is a feature whose **Feature F-measure** is **maximized** by the group members (i.e. data).

Quality based on data description space

Feature maximization founding principle [Lamirel 04]



The R, P, F–measure criteria are independent of the grouping method (symbolic equivalence).

Feature maximization metric

Extended use

- Local values of unsupervised Precision and Recall quality indexes can be used for efficiently extracting association rules [Lamirel 10]
- In machine learning feature maximization metric proved to have very various use, like:
  - Optimizing learning [Attik 06]
  - Cluster labeling and cluster content mining [Lamirel 08]
  - Detecting incoherent clustering results [Lamirel 10]
  - Substituting to distance in clustering [Lamirel 11][Lamirel 12]
  - Efficient feature selection for supervised classification [Lamirel 11]
  - SNA analysis and data synthesis and summarization [Ongoing]
  - Setting up new cluster quality indexes [Here]

Adaptation of feature maximization metric to feature selection

The feature maximization process can be applied on classes as well as on clusters as soon as it is only depending on associated data. It is a parameter-free process.

The set S_c of features that are characteristic of a given class c belonging to an overall class set C results in:

$$S_c = \{ f \in F_C \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \}$$

where $\overline{FF}(f) = \sum_{c' \in C} FF_{c'}(f) / |C_{/f}|$ and $\overline{FF}_D = \sum_{f \in F} \overline{FF}(f) / |F|$.

and $C_{/f}$ represents the restriction of the set C to the classes in which the feature f is represented.

Finally, the set of all the selected features S_C is the subset of F defined as:

$$S_C = \bigcup_{c \in C} S_c$$

Additional F-max metric based contrasting (FMC)

Contrast or information gain characterizes the strength of the relation between a feature and a class. For a feature *f* associated to a class *c*, it can be expressed as:

$$C_c(f) = (FF_c(f)/\overline{FF}(f))^k$$

- A contrast value > 1 an active behavior of the feature in the class,
- A contrast value < 1 an passive behavior of the feature in the class,
- The magnification factor *k* is used to enhance contrast in a non linear way for facilitating class separation.

A simple example

- We consider a sample of Men (M) and Women (F) for which we measure Hair_length and Shoes_size and Nose_size :

Shoes _size	Hair _length	Nose _size	Class
9	5	5	M
9	10	5	M
5	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

A simple example

- We compute the Feature Recall (FR) and the Feature Precision (FP) and the Feature F-measure (FF) for each class and each feature and each class:

Shoes _size	Hair _length	Nose _size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

FR(S,M) = 27/43 = 0.62

FP(S,M) = 27/78 = 0.35

$$FF(S,M) = \frac{2(FR(S,M) \times FP(S,M))}{FR(S,M) + FP(S,M)} = 0.48$$

A simple example

- We compute the average marginal values of Feature F-measure by feature (local) and the overall Feature F-measure for each class and each feature and each class:

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Hair_ length	0.39	0.66	0.53
Shoes_ size	0.48	0.22	0.35
Nose_ size	0,3	0,24	0,27

$\overline{F(.,.)}$
0.38

The features whose Feature F-measure is under the global Feature F-mesure average are removed

⇒ Nose_size is removed

The remaining (i.e. selected) features whose F-measure is over marginal average in one class are considered as active in this class

=> **Shoes_size is active in Men class**

=> Hair_length is active in Women class

A simple example

- The contrast factor highlights the degree of activity/passivity of selected features relatively to their marginal Feature F-measure average in the different classes:

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Hair _length	0.39	0.66	0.53
Shoes _size	0.48	0.22	0.35

	C(x,M)	C(x,F)
Hair _length	0.39/0.53	0.66/0.53
Shoes _size	0.48/0.35	0.22/0.35

The constrast can be seen as a function that will tend to:

1. Overlength the Hairs of Women
2. Oversize the Shoes of Men
3. Underlength the Hairs of Men
4. Undersize the Shoes of Women

	C(x,M)	C(x,F)
Hair _length	0.74	1.25
Shoes _size	1.37	0.63

A simple example

- The contrast is applied on the data in order to modify the feature weights depending on the data class:

Shoes _size	Hair _length	Class
9	5	M
9	10	M
9	20	M
5	15	F
6	25	F
5	25	F

Original data

Shoes _size	Hair _length	Class
12,33	3.7	M
12,33	7.4	M
12,33	14.8	M
3.15	18.75	F
3,78	31.25	F
3.15	31.25	F

Contrasted data

Data contrast can change the organization of the data in the description space in a non linear way.

A simple example

- The magnification factor (k) can enhance the contrast to facilitate classification in complex cases:

Shoes_size	Hair_length	Class
12,33	3.7	M
12,33	7.4	M
12,33	14.8	M
3.15	18.75	F
3,78	31.25	F
3.15	31.25	F

Contrasted data (k = 1)

Shoes_size	Hair_length	Class
28.30	1.59	M
28.30	3.19	M
28.30	6.37	M
0.99	36.47	F
1.20	60.79	F
0.99	60.79	F

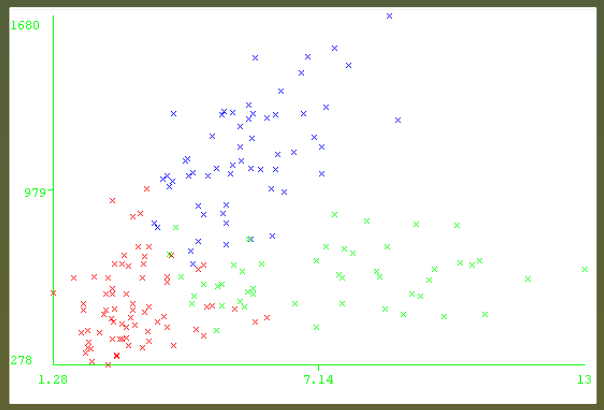
Contrasted data (k = 4)

Magnification is a non linear transformation (enhance non-linearity).

A simple real case

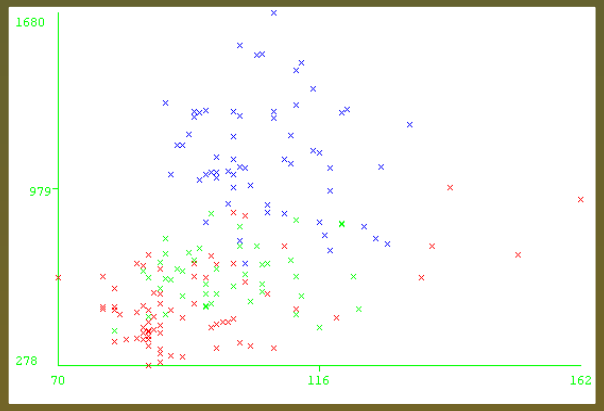
Wine dataset with J48 or FMC

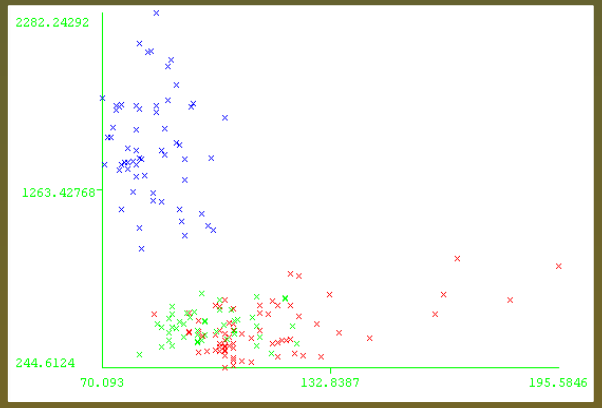
J48

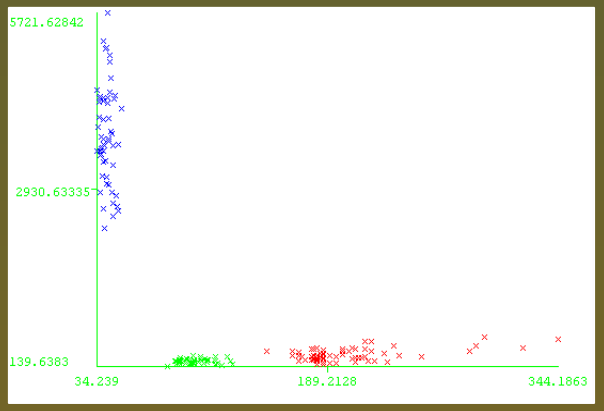


J48 and FMC select both 2 features among 13 but discrimination become better with FMC when magnification factor is increased

FMC







k = 1 k = 2 k = 4

Classification with FMC

Deft challenge [JADT 2014]

- Dataset of extracts of talk of CHIRAC et MITTERAND presidents:
  - 73255 sentences of Chirac
  - 12320 sentences of Mitterrand,
- Best results till now on that dataset : 88% accuracy (almost 16850 bilateral errors) by LIA,
- Result with feature maximization : 99,999% accuracy (12 unilateral errors)
  - Extra-light NLP preprocessing
  - No lemmatization is needed
  - Stop words are kept and proof to be useful for analysis,
- Feature selection reduced the description space from ≈ 50000 to ≈ 5000 dimensions.

CHIRAC	MITTERAND
1.930810 partenariat	1.881835 douze
1.858265 dynamisme	1.852007 est-ce
1.811123 exigence	1.800091 eh
1.775048 compatriotes	1.786760 quoi
1.769069 vision	1.777568 -
1.768280 honneur	1.758319 gens
1.763166 asie	1.747909 assez
1.762665 efficacité	1.741650 capables
1.745192 saluer	1.716491 penser
1.743871 soutien	1.700678 bref
1.737269 renforcer	1.688314 puisque
1.715155 concitoyens	1.672872 on
1.709736 réforme	1.662164 étais
1.703412 devons	1.620722 parle
1.695359 engagement	1.618184 fallait
1.689079 estime	1.604095 simplement
1.671255 titre	1.589586 entendu
1.669899 pleinement	1.580018 suite
1.662398 cœur	1.572140 peut-être
1.661476 ambition	1.571393 espère
1.654876 santé	1.560364 parlé
1.640298 stabilité	1.550856 dis
1.632421 amitié	1.549594 cela
1.628630 accueil	1.538523 existe
1.622473 publics	1.535598 façon
1.616558 diversité	1.529225 pourrait
1.614945 service	1.525645 là
1.612488 valeurs	1.525508 chose
1.610123 détermination	1.523575 époque
1.601097 réformes	1.522290 production
1.592938 état	1.519365 trouve
……..	……..

Quality evaluation using feature maximization (Principle)

- A good clustering model should be able to maximize the weighted sum of positive contrast in clusters (≈ generic intra-cluster inertia):

$$PC_k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|S_i|}\sum_{f\in S_i} G_i(f)$$

The intuition behind this approach is that active features are relevant on their own of clusters inherent structure and maximal averaged contrast on that features is directly related with the most relevant global clustering structure.

Quality evaluation using feature maximization (Principle) (2)

- A more complete approach could combine weighted summation of positive and weighted summation of invert of negative contrast

  (≈ generic intra-cluster and inter-cluster inertia):

$$EC_k = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{|s_i|\sum_{f\in S_i} G_i(f) + |\overline{s_i}|\sum_{h\in \overline{S_i}} \frac{1}{G_i(h)}}{|s_i| + |\overline{s_i}|}\right)$$

  The intuition behind this approach is that passive features plays also an important role for highlighting optimal model and that optimal global clustering model is the model with the highest structural gaps.

Quality evaluation using feature maximization (Principle) (3)

- CB index is a combination of the 2 other approaches:

Algorithm 1 CB : Combining PC and EC indexes.

▷ $PC(i)$ returns the PC value of model i,
▷ $EC(j)$ returns the EC value of model i,
▷ $Peak(F(i))$ returns true if $F(i-1) < F(i) > F(i+i)$, for i $\in \{2, \cdots, k-1\}$.

```
procedure CB(List L of 1..k models)
    sort(L) by decreasing order according to (EC + PC) value
    for i in L do
        if Peak(PC(i)) then
            return i;
        end if
    end for
    return −out−;
end procedure
```

Quality evaluation using feature maximization (Evaluation process)

- Evaluation must be conducted on dataset of various dimension and size:

	IRIS	IRIS-B	WINE	PEN	ZOO	VRBF	R8	R52
Nb. class	3	3	3	10	7	12-16	8	52
Nb. data	150	150	178	10992	101	2183	7674	9100
Nb. feat.	4	12	13	16	114	231	3497	7369

- Purity value is used in a complementary way to take into account sub-optimal results generally produced by clustering (as compared as ground truth):

$$mPur = \frac{\sum_{c \in C, |prev(c)|>1} |prev(c) \cap c|}{n}$$

Quality evaluation using feature maximization (Evaluation process)

- Several clustering methods are used:
  - K-means [MacQueen 67]
  - GNG [Fritske 95]
  - IGNGF [Lamirel 11] (proven to outperform other methods on binary or frequency data),
- The size of the clustering model is let varying from unity till 3 times the ground truth,
- The estimation produced by a index is considered as valid in the range between the ground truth and the maximal value or interval of purity (MaxP),
- Increasing amount of noise is introduced in clustering results to test the stability of the quality indexes.

Quality evaluation using feature maximization (Results – low dimension)

	IRIS	IRIS-B	WINE	PEN	SOY	Number correct
DB	2	5	5	7	19	1/5
CH	2	3	6	8	4	1/5
DU	1	1	8	17	8	0/5
SI	4	2	7	14	4	0/5
XB	2	7	-out-	19	-out-	0/5
AIC	2	4	2	14	24	0/5
BIC	-out-	4	-out-	14	-out-	0/5
NEG	9	ND	-out-	14	ND	0/5
PC	3	3	4	14	8	3/5
EC	6	3	4	11	4	3/5
CB	3	3	4	11	10	4/5
MaxP	3	3	3-4	12	18-20	
Truth	3	3	3	10	16	
Method	K-means	K-means GNG	K-means	K-means	GNG	

PC, EC, and more especially CB, clearly outperform other approaches in realistic optimal clustering model evaluation for low dimensional problem.

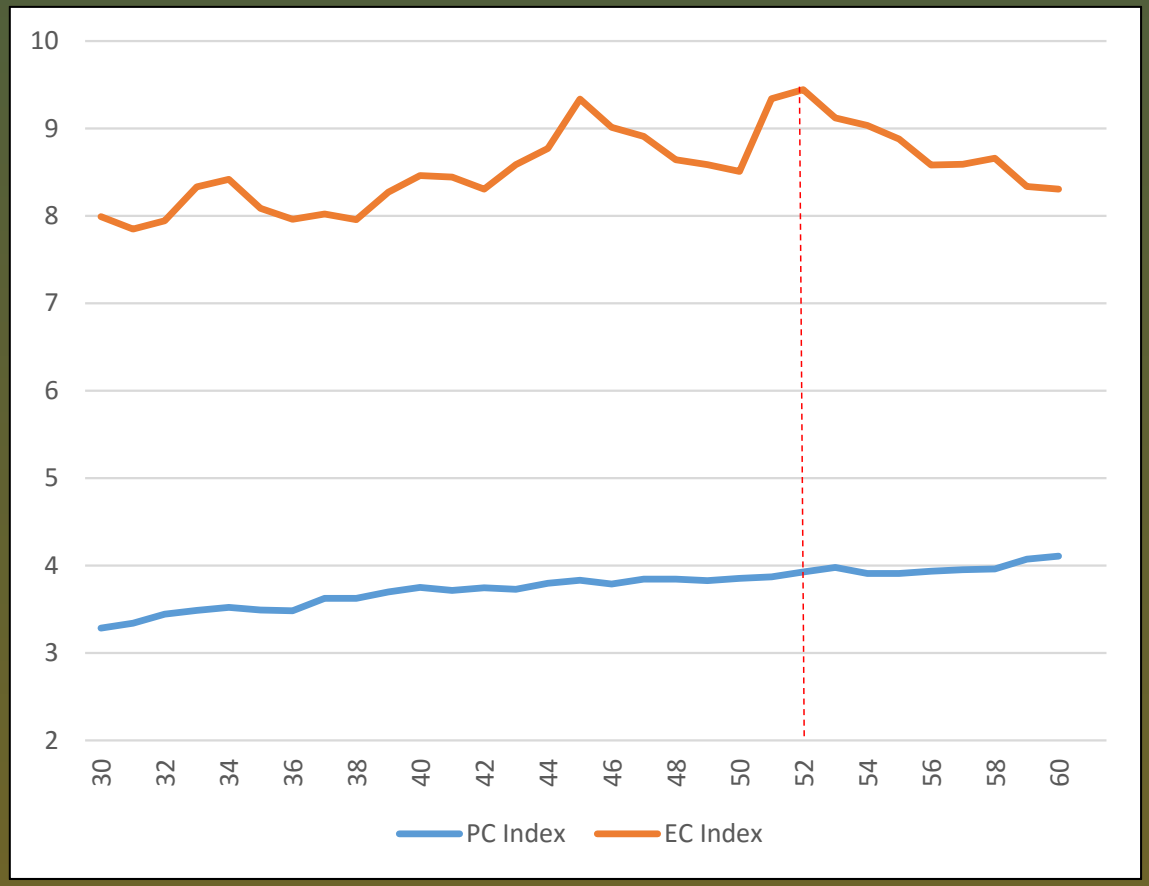
Quality evaluation using feature maximization (Results – high dimension)

	ZOO	VRBF	R8	R52	Number correct
DB	8	-out-	5	58	2/4
CH	4	7	6	-out-	0/4
DU	8	2	-out-	-out-	1/4
SI	4	-out	-out-	54	1/4
XB	-out-	23	-out-	-out-	0/4
AIC	2	14	3	30	1/4
BIC	-out-	-out-	-out-	58	1/4
NEG	ND	ND	ND	ND	0/4
PC	6	29	-out-	-out-	0/4
EC	7	17	6	52	3/4
CB	7	17	13	53	4/4
MaxP	10	12-17	13	54-58	
Truth	7	12-16	8	52	
Method	IGNGF	GNG IGNGF	GNG IGNGF	IGNGF	

EC, and more especially CB, still clearly outperform other appro
optimal clustering model evaluation for high dimensional proble

Computation time is low: EC = 125s – SI = 43000s on R52

Quality evaluation using feature maximization (Index divergence case)



An index is divergent if it does not find the optimal model in a reasonable range around ground truth.

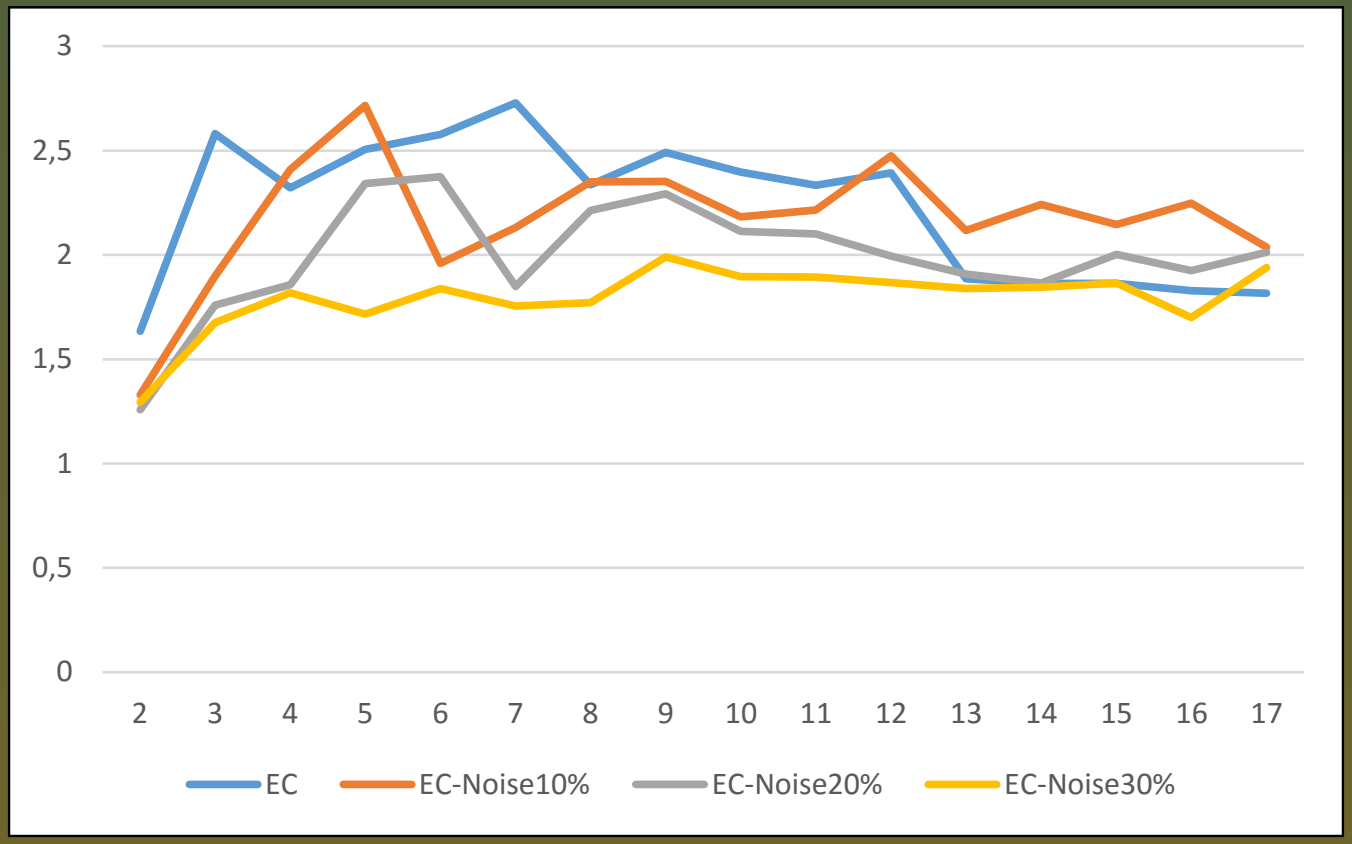
Quality evaluation using feature maximization (Resistance to additional noise)

- Data of clusters are migrated in a random way to other clusters to different fixed amount for all model sizes,
- Indexes are recalculated on noised models to look for potential variation in their behavior,
- This experiment highlights robustness to weak clustering results.

	ZOO	ZOO Noise 10%	ZOO Noise 20%	ZOO Noise 30%	Number of correct matches
DB	8	4	3	3	1/4
CH	4	5	3	3	0/4
DU	8	2	2	2	1/4
SI	4	-out-	-out-	-out-	0/4
XB	-out-	-out-	-out-	-out-	0/4
PC	6	4	11	9	1/4
EC	7	5	6	9	2/4
CB	7	5	6	9	2/4
MaxP	12	10	10	12	
Method	IGNGF	IGNGF	IGNGF	IGNGF	

EC, and more especially PC, never get “out of work” even when noise is increasing to a significant extent.

Quality evaluation using feature maximization (Resistance to noise)



PC index behavior is smoothing (i.e. degrading) progressively with noise.

Quality evaluation using feature maximization

Real life case - Research Topic Changes Tracking (RTCT)

ISTEX Data

- ISTEX data are issued from different scientific editors, and there is no standardization of metadata or even no available metadata in some cases
- The exploited method must be able to tackle with large collection in an unsupervised way
  (time efficiency + a few of even no parameters)
- Overall time period lengths including stable topics can vary over time
- Visualization of diachronic changes is still on open problem

A first subset of ~10000 papers related to health care is extracted to perform a feasibility study on diachronic analysis

Quality evaluation using feature maximization

Real-life case – RTCT – Overall methodology

General query

Sub collection

Indexing (Metadata)

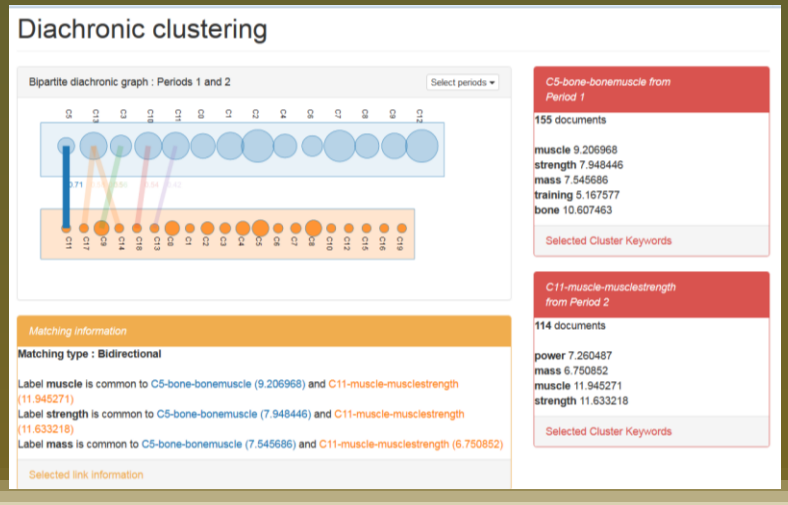
ISTEX repository

Automatic period detection

Clustering + Optimal model detection

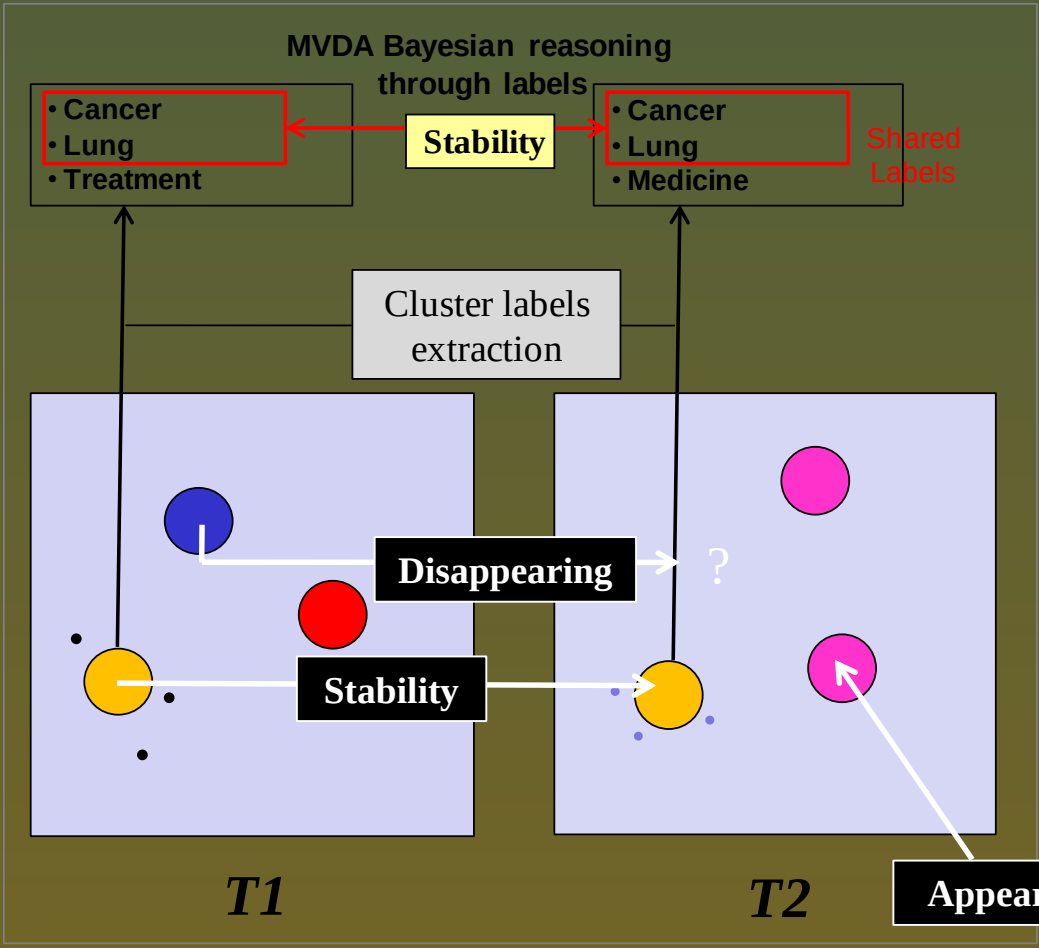
Diachronic analysis

Visualization (Diachronic'Explorer)



Quality evaluation using feature maximization

Real-life case – RTCT – Topic matching principle



Multiple functions of the MVDA model are exploited :

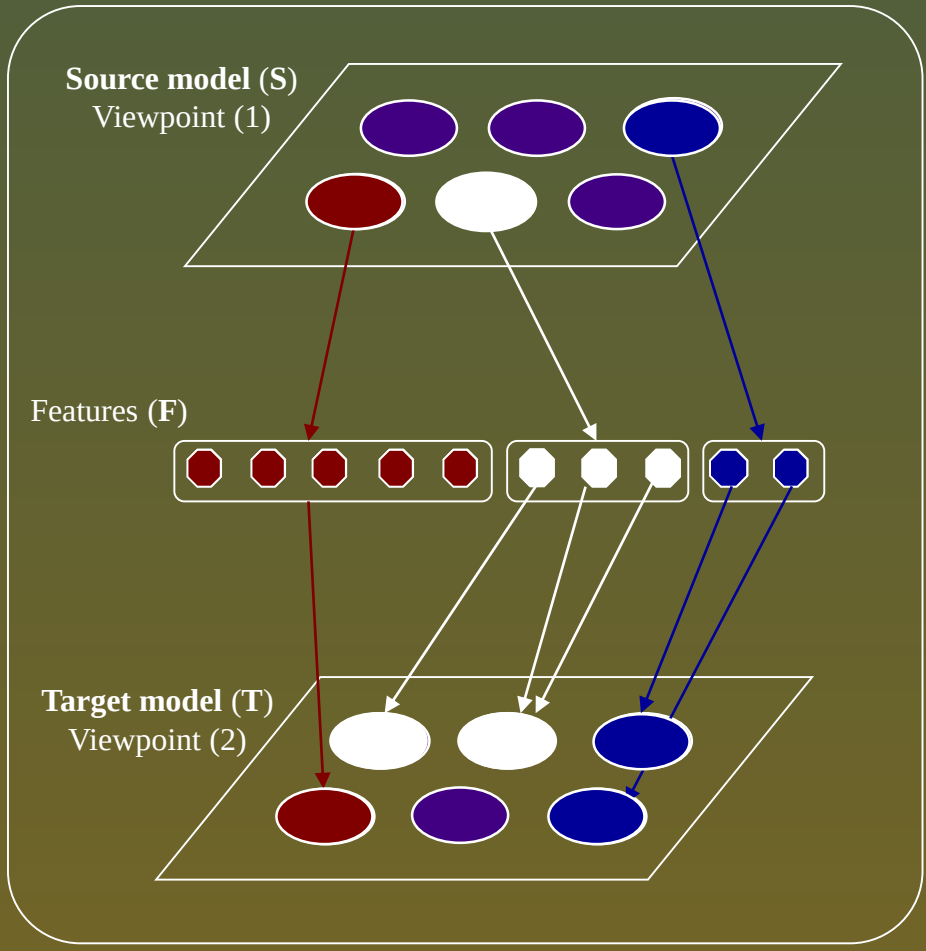
- Time subperiods associated to viewpoints
- Optimized clustering using neural methods and unbiased quality measures
- High performance (F-max based) labeling techniques
- Adapted Bayesian reasoning on labels

F-max cluster labeling provides dimensionality reduction (feature selection) .

Quality evaluation using feature maximization

Real-life case – RTCT – MVDA Bayesian reasoning

- MVDA paradigm relies on Bayesian reasoning
- Bayesian network is generated in an unsupervised way
- Uses clustering models shared data to perform cluster comparisons
- Applicable with any clustering method



Bayesian network model :

$$P(act_m | T_j, Q) = \frac{\sum_{d \in act_m, T_j} Sim(d, S_i)}{\sum_{d \in T_j} Sim(d, S_i)}$$

Quality evaluation using feature maximization

Real-life case – RTCT – Topic matching principle

Comparison is performed using an adaptation of MVDA Bayesian reasoning with :

$$P(t|s) = \frac{\sum_{l \in L_s \cap L_t} L - F_t(l)}{\sum_{l \in L_t} L - F_t(l)}$$

where L_x represent the set of labels associated to the cluster x, and $L_x \cap L_y$ represent the common labels, which can be called the label **matching kernel**, between the cluster x and the cluster y.

Quality evaluation using feature maximization

Real life case – RTCT – Topic matching principle

The **similarity** between a cluster s of the source period and a cluster t of the target period is established using :

- The average matching probabilities $P_A(x)$ of a period cluster
- The global average activity A_x generated by a period model on the model of the alternative period and its standard deviation $б_x$

Similarity is found if :

1) $P(t \mid s) > P_A(s)$ *et* $P(t \mid s) > A_s + б_s$

2) $P(s \mid t) > P_A(t)$ *et* $P(s \mid t) > A_t + б_t$

Cluster splitting, cluster merging, vanishing clusters, appearing clusters events can be deduced from former similarity rules.

Quality evaluation using feature maximization

Real-life case – RTCT – Example of results

Theory to practice

source cl [12/7]
- Stable
f1: 0.259 s (***)
f1: 0.086864[23] f2: 0.129486[2] conducting polymers (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.034510[23] f2: 0.000000[-1] Experimental study

- Highly dominant (or peculiar) labels in target period
f1: 0.072006[23] f2: 0.206426[2] Polymer films (***)
f1: 0.054435[23] f2: 0.114637[2] Polymer blends (***)
f1: 0.000000[-1] f2: 0.039558[2] Spin-on coating
f1: 0.000000[-1] f2: 0.028204[2] Polymerization

New component

source clus 24 [20/8]
- Stable la
f1: 0.03837 pound (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.043265[15] f2: 0.000000[-1] MIS structure
f1: 0.026522[15] f2: 0.000000[-1] Diamond

- Highly dominant (or peculiar) labels in target period
f1: 0.061132[15] f2: 0.222402[24] Amorphous semiconductors (***)
f1: 0.054647[15] f2: 0.131473[24] Hydrogen (***)
f1: 0.000000[-1] f2: 0.067403[24] Selenium
f1: 0.000000[-1] f2: 0.039028[24] Plasma CVD coatings

Theory to practice

source clust [29/7]
- Stable lab
f1: 0.03572 laser (***)

- Highly dominant (or peculiar) labels in source period
f1: 0.148633[14] f2: 0.057783[14] Semiconductor laser (***)
f1: 0.078080[14] f2: 0.033436[14] Laser diodes (***)
f1: 0.026498[14] f2: 0.000000[-1] Surface
f1: 0.026027[14] f2: 0.000000[-1] Waveguide laser

- Highly dominant (or peculiar) labels in target period
f1: 0.000000[-1] f2: 0.068895[14] Light sources
f1: 0.000000[-1] f2: 0.039487[14] Laser beam applications
f1: 0.000000[-1] f2: 0.029637[14] Vertical cavity laser
f1: 0.000000[-1] f2: 0.025024[14] VCSEL

Vocabulary change

source clust 7/13]
- No stable

- Highly dominant (or peculiar) labels in source period
f1: 0.266901[24] f2: 0.068167[33] Optical fabrication (***)
f1: 0.045998[24] f2: 0.000000[-1] Integrated circuit technology
f1: 0.042258[24] f2: 0.000000[-1] Interference filter
f1: 0.041773[24] f2: 0.000000[-1] Semiconductor technology

- Highly dominant (or peculiar) labels in target period
f1: 0.077799[24] f2: 0.213749[33] Optical design techniques (***)
f1: 0.000000[-1] f2: 0.055834[33] Aberrations
f1: 0.000000[-1] f2: 0.039636[33] Ray tracing

source cluster 16 is vanishing
f1: 0.141849[16] f2: 0.000000[-1] Optical fiber
f1: 0.078762[16] f2: 0.000000[-1] Fiber laser
f1: 0.060706[16] f2: 0.000000[-1] Acoustooptical device
f1: 0.049628[16] f2: 0.000000[-1] Ring laser

target cluster 9 is appearing
f1: 0.035520[5] f2: 0.160462[9] Fluorescence
f1: 0.000000[-1] f2: 0.082686[9] Phosphorescence
f1: 0.063888[1] f2: 0.105132[9] Exciton

target cluster 39 is appearing
f1: 0.000000[-1] f2: 0.144184[39] Pixel
f1: 0.000000[-1] f2: 0.110076[39] CMOS image sensors
f1: 0.000000[-1] f2: 0.077578[39] Chip
f1: 0.000000[-1] f2: 0.060044[39] High sensitivity

Quality evaluation using feature maximization

Real life case – RTCT – Quality evaluation

Criteria 1 : number of matches

Criteria 2 :

$$QMA = \sum_{i,j \in M} |S_{ij}| * \frac{P(i|j) + P(j|i)}{2}$$

where M represents the set of couple of clusters for which a match is detected.

In temporal matching process, hypothesis is that an accurate model selection will produce the larger number of matches, with matching kernels of the largest sizes and with the highest matching probability.

We consequently exploit two complementary criteria for the evaluation of the behavior of the indexes.

Quality evaluation using feature maximization

Real life case – RTCT – Quality evaluation results

	Opt. Period P1	Opt. Period P3	Opt Period P3	Number of temporal matches	QMA evaluation criteria
DB	-out-	-out-	-out-	0	0
CH	3	4	4	5	15.26
DU	14	20	-out-	6	8.60
SI	-out-	-out-	-out-	0	0
XB	-out-	-out-	-out-	0	0
PC	23	323	-out-	9	10.61
EC	10	6	8	13	27.20

Temporal matching results confirm the better performance of EC index as compared to other indexes.

Summary of the results

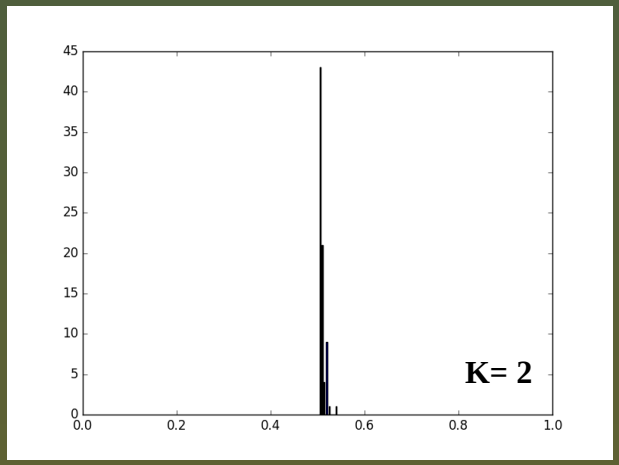
	Behaviour in low dimension	Behaviour In high dimension	Resistance to noise	Diachrony	Independance To distance measures
DB	-	+/-	+/-	--	No
CH	-	-	+/-	+	No
DU	--	-	-	+/-	No
SI	--	-	--	--	No
XB	--	--	NE	NE	No
AIC	-	-	NE	NE	No
BIC	--	--	NE	NE	No
NEG	--	--	NE	NE	Yes
SBS	NE	NE	NE	NE	Yes
PC	+	+	+	+	Yes
EC	+	++	++	+ +	Yes
CB	+ +	++	++	+ +	Yes

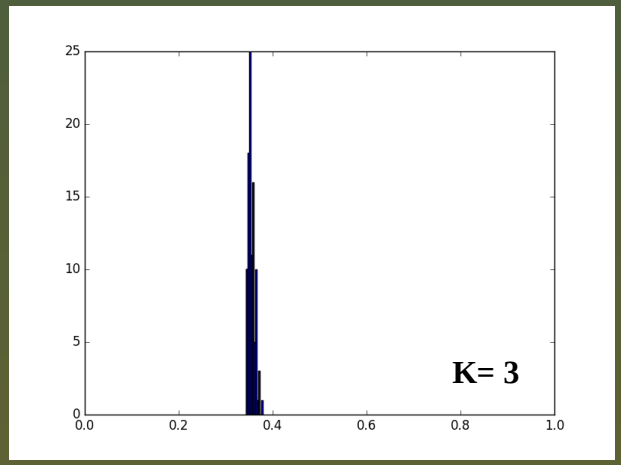
Conclusion and perspectives

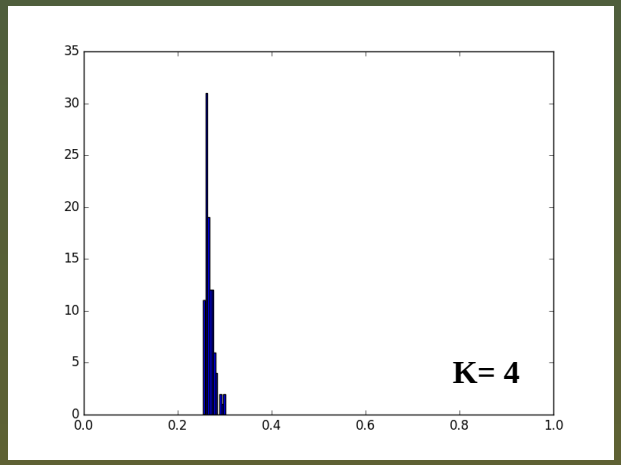
- We have proposed new clustering quality indexes based on feature maximization metric,
- Feature maximization metric is based on a cross-domain approach (numeric + symbolic + IR),
- Method aims at finding the model that maximize the information carried by most representative features of the model (instead of using error),
- Method outperforms usual indexes as well as our former proto-indexes on high dimensional context,
- Proposed method is accurately resisting to noise naturally present in clustering,
- Methods works in usual test sets as well as in real-life applications like for temporal matching,
- Computation cost is low,
- Adaptation to fuzzy clustering is straightforward.

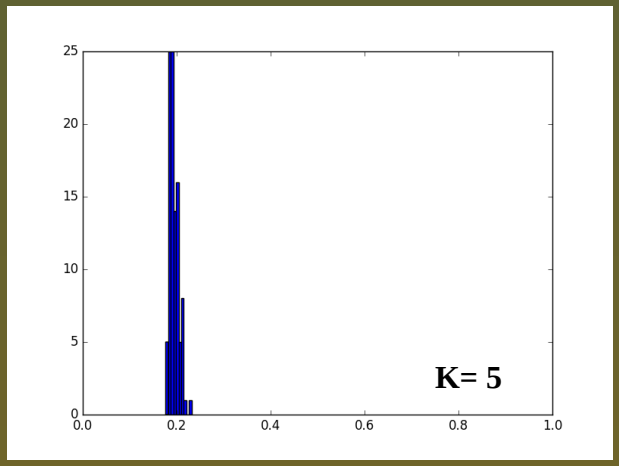
Subsampling method

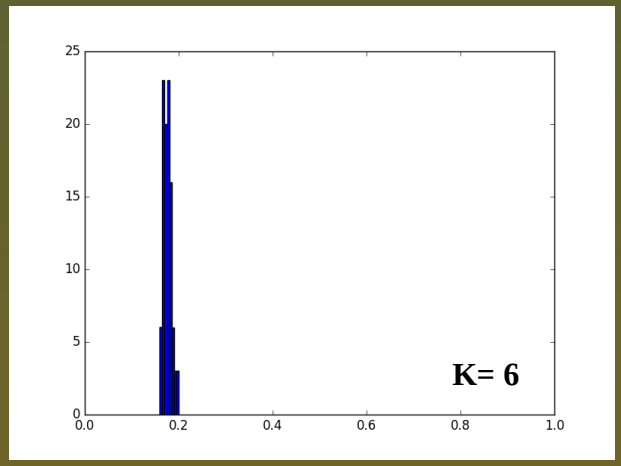
Results for Iris

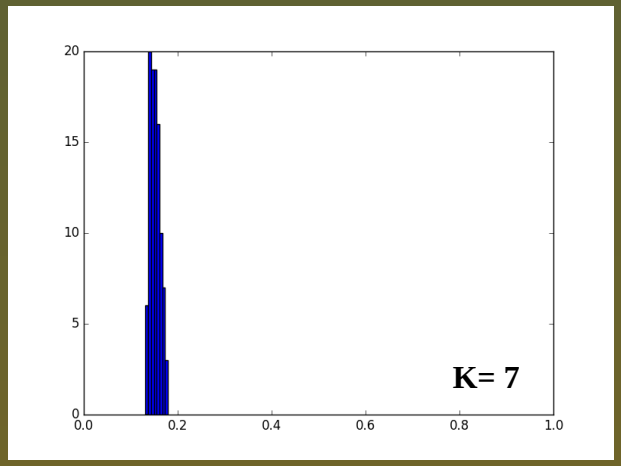








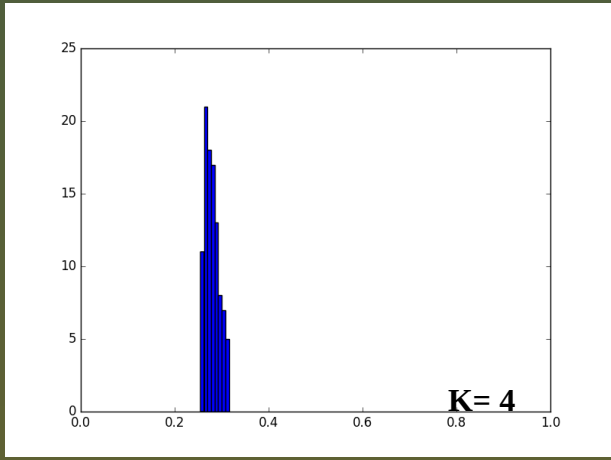


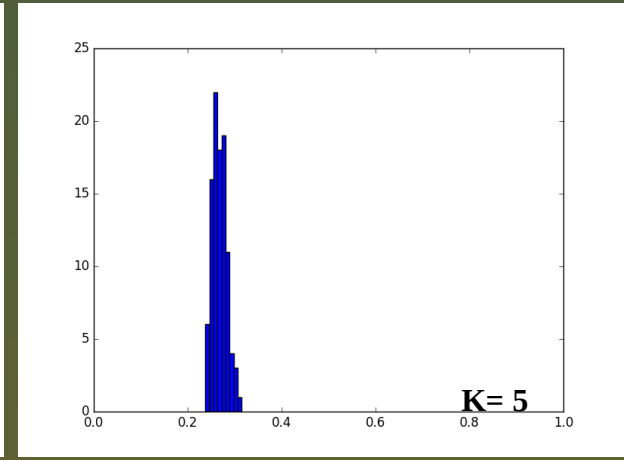


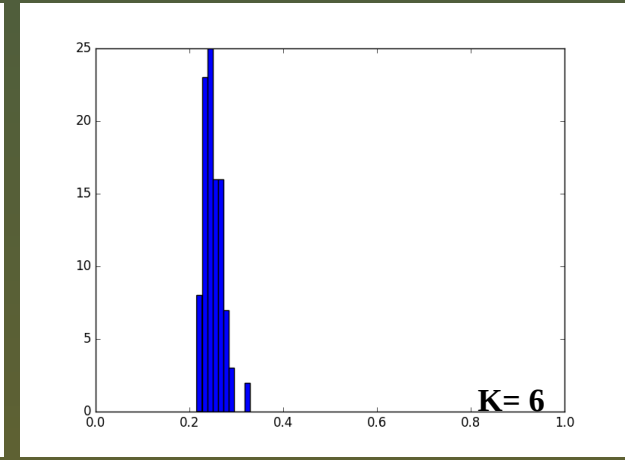
No decision on correlation distribution drastic decrease can be taken (might be due to suboptimal results of clustering)

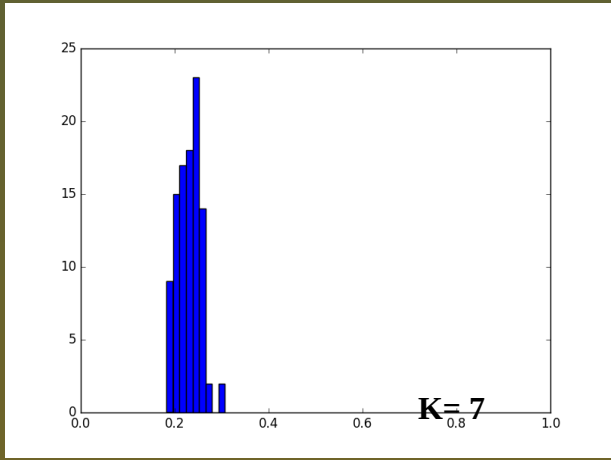
Subsampling method

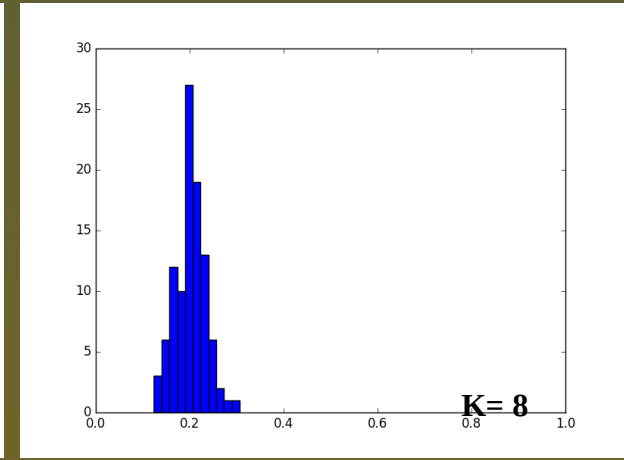
Results for Zoo

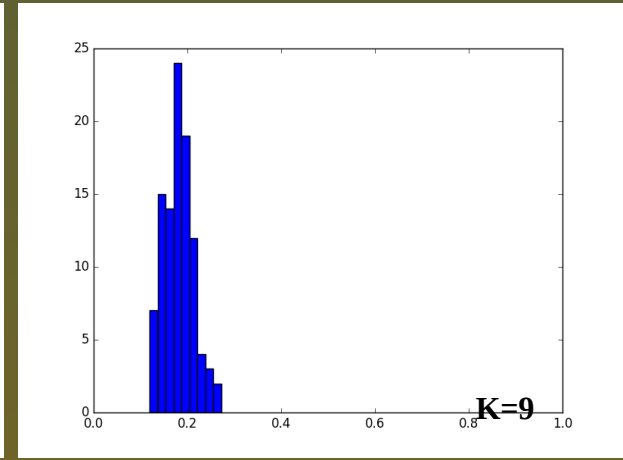








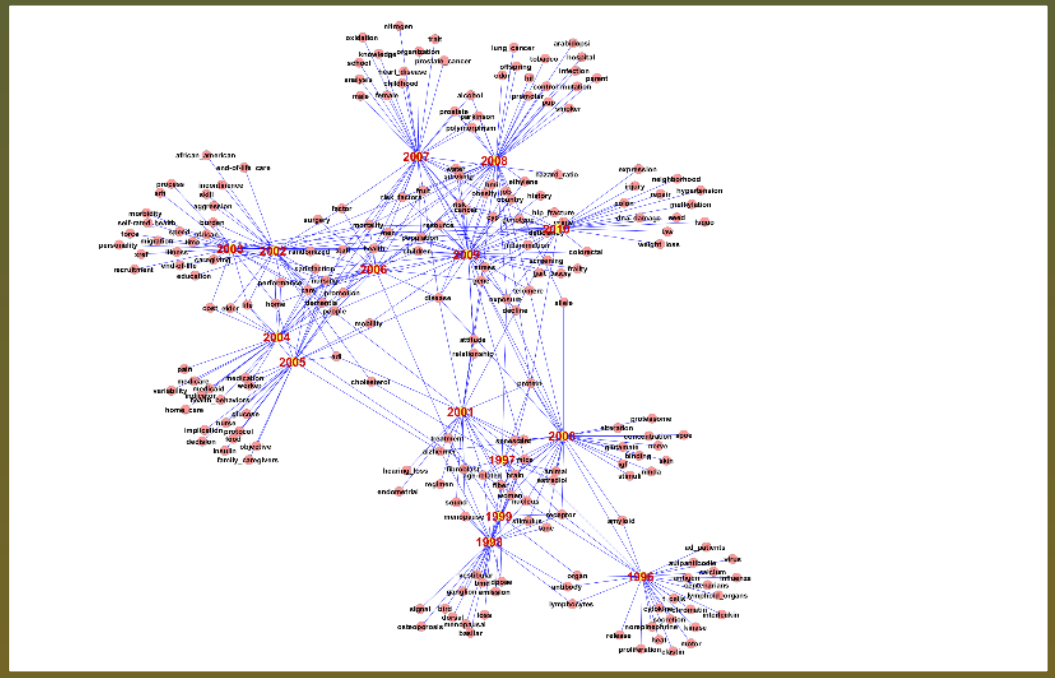




No decision on correlation distribution drastic decrease can be taken (might be due to suboptimal results of clustering)

Conclusion and perspectives (2)

- We have to perform larger scope experiments including more indexes and more especially entropy-based indexes,
- We plan to experience a new approach based on the analysis of properties of contrast graphs.



Contact and questions

email: lamirel@loria.fr

Github with codes of the quality indexes and test data:

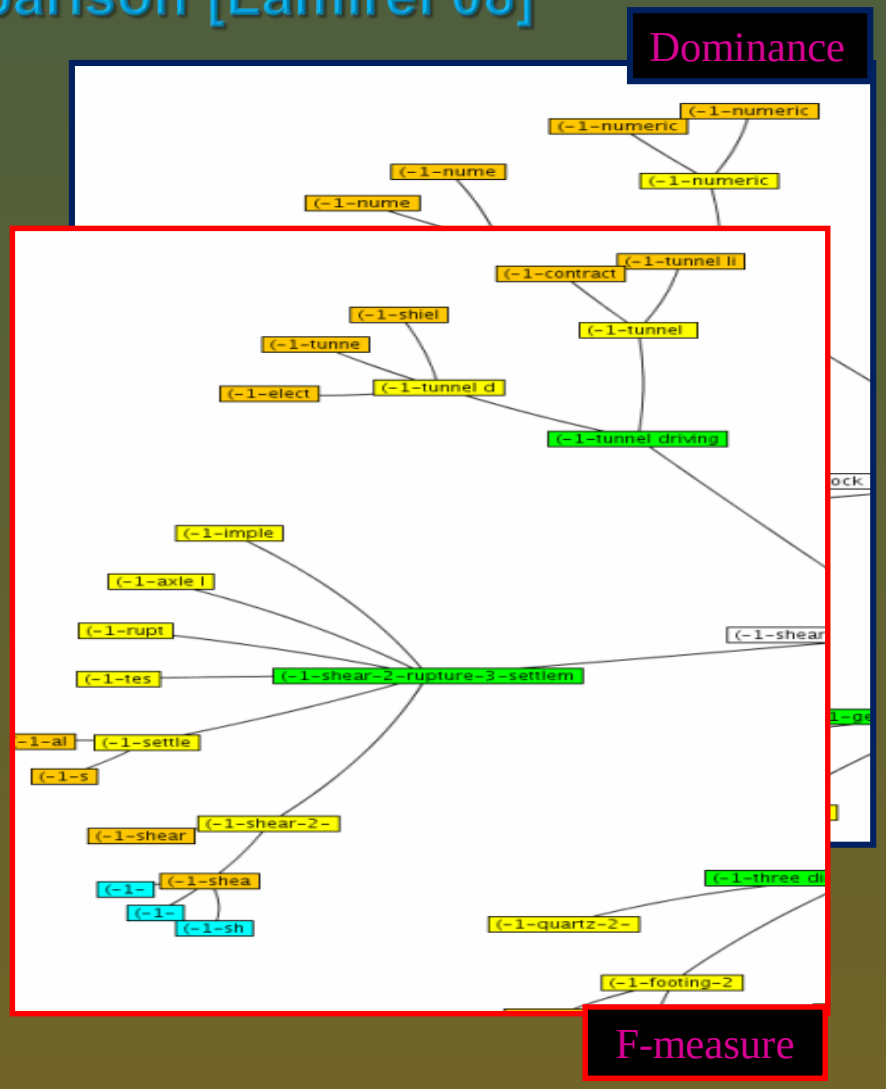
http://github.com/nicolasdugue/clusteringQuality

Cluster labeling based on feature maximization

Labeling method comparison [Lamirel 08]

	PLS	AVP	PSS	PHL
Dominance	1216	0.03	568	525
Frequency	245	0.24	166	592
F-Measure	155	0.26	112	760
Chi²	121	0.21	89	1485

PLS : Penalty of Leave Similarity,
ALP : Average Leave labels Precision,
PSS : Penalty of Sons Similarity,
PHL : Penalty of Labeling Heterogeneity.



F-measure provides the best compromise: exhaustivity – discriminance [Lamirel 08].