Prototype-Based Classification of Graphs

Romain Raveaux[1], Eugen Barbu[2], Sébastien Adam[3]
Pierre Héroux[3], Eric Trupin[3]

[1]L3I – Université de La Rochelle

[2]ACM Professional Member

[3]LITIS – Université de Rouen

July 23-07, 2009

Content

The underlying question

Background

- The selection of graph prototypes.
- In a context of supervised classification.
- Using structured data.

The underlying question.

- How to find graph prototypes for a classification purpose ?

The supervised classification of directed labeled graphs

Graph notations

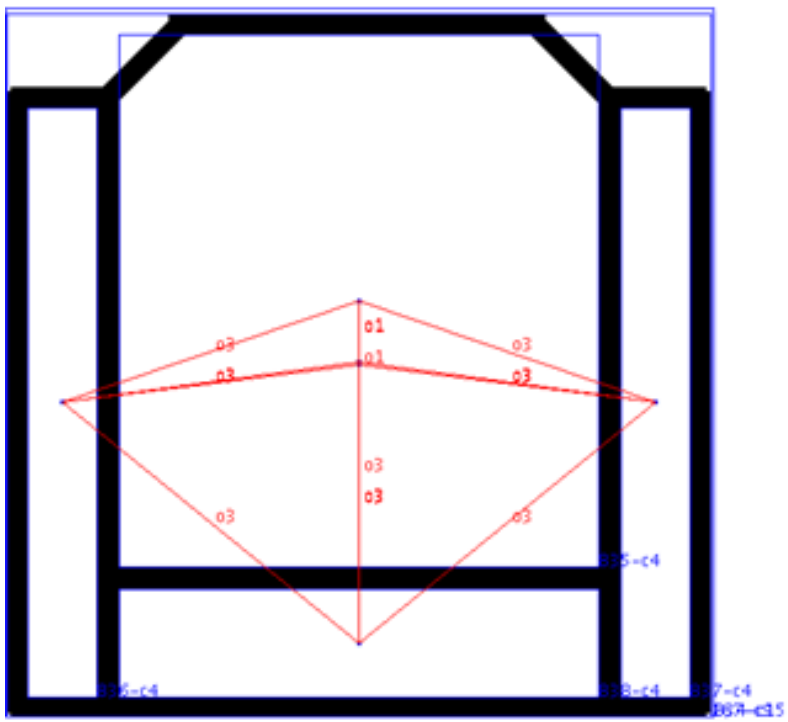
Let L_V and L_E denote the set of node and edge labels, respectively. A labeled graph G is a 4-tuple $G = (V, E, \mu, \xi)$, where

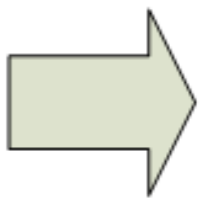
- V is the set of nodes,
- $E \subseteq V \times V$ is the set of edges
- $\mu : V \rightarrow L_V$ is a function assigning labels to the nodes, and
- $\xi : E \rightarrow L_E$ is a function assigning labels to the edges.

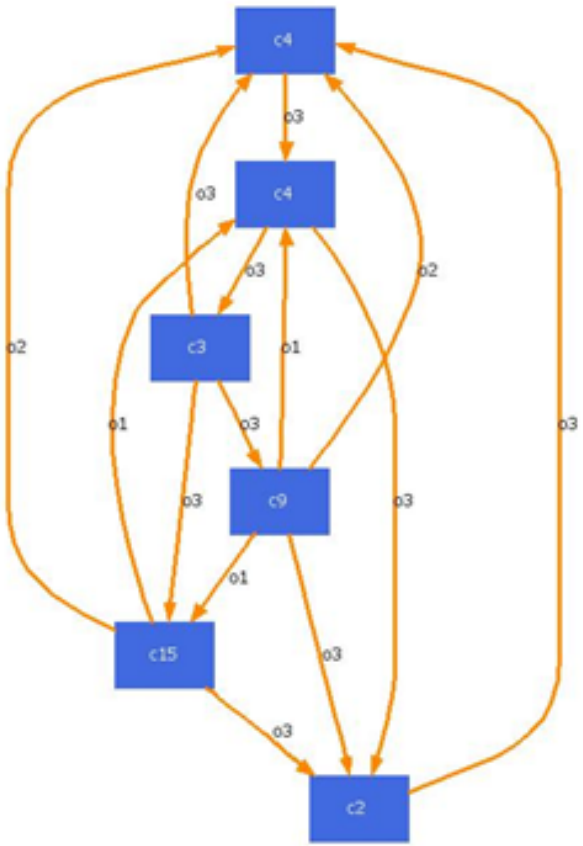
Classification

- Training set : $X_{tr} = \{x_1, .., x_L\}_{l=1}^{L}$ with N classes.
- Vectors of prototypes : $V = \{v_{1\,1}, .., v_{n\,m}\}$ with M prototypes per class et $|V| = M \times N$.
- Error on the test set: $E_{nn}(X_{test}; V)$, Learning of meta-parameters.
- Error on the validation base : $E_{nn}(X_{val}; V)$, Final results.

An example: From symbol to graph



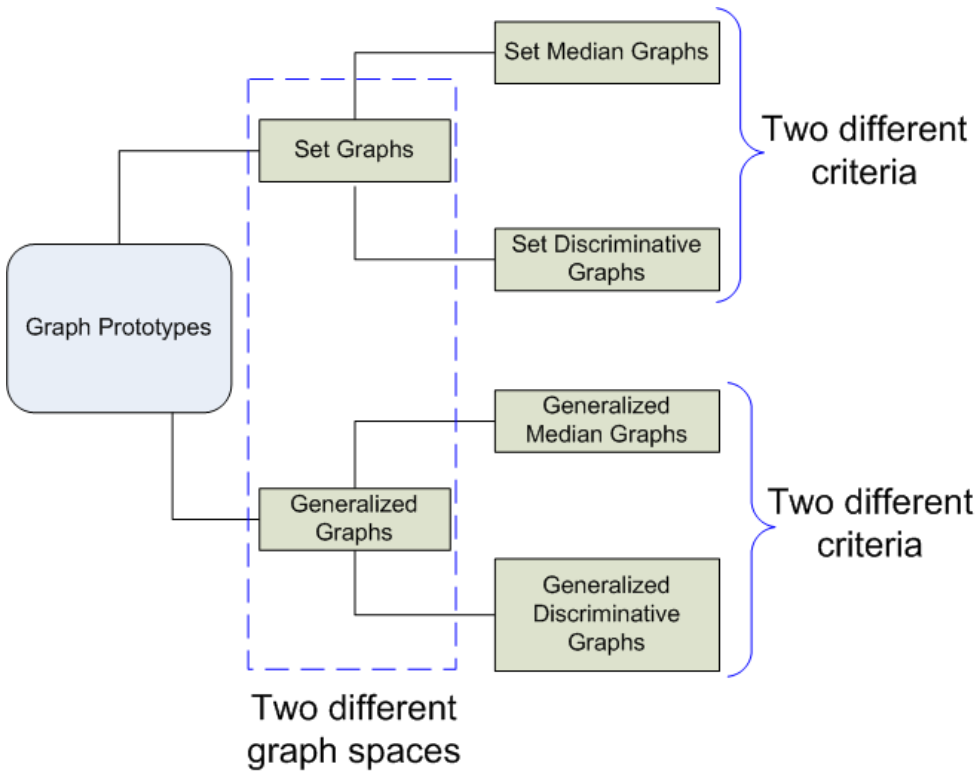




The nature of graph prototypes

Graph prototypes nature stands apart in two ways :

- The space the prototype belongs to.
- The criteria in use to reach the prototype.



Median Graphs

- The criterion → minimization of the Sum Of Distances (SOD) for a given class.
- Set Median Graphs.
- Generalized Median Graphs.

Definition

Set Median Graph (smg)

$$\widehat{smg} = arg \min_{g \in S} \sum_{i=1}^{n} d(g, g_i)$$

Definition

Generalized Median Graphs (gmg)

$$\overline{gmg} = arg \min_{g \in U} \sum_{i=1}^{n} d(g, g_i)$$

Discriminative Graphs

- The Criterion $\longrightarrow$ minimization of the error of classification on a test set.
- Set Discriminative Graphs.
- Generalized Discriminative Graphs.

Definition

Set Discriminative Graphs (sdg)

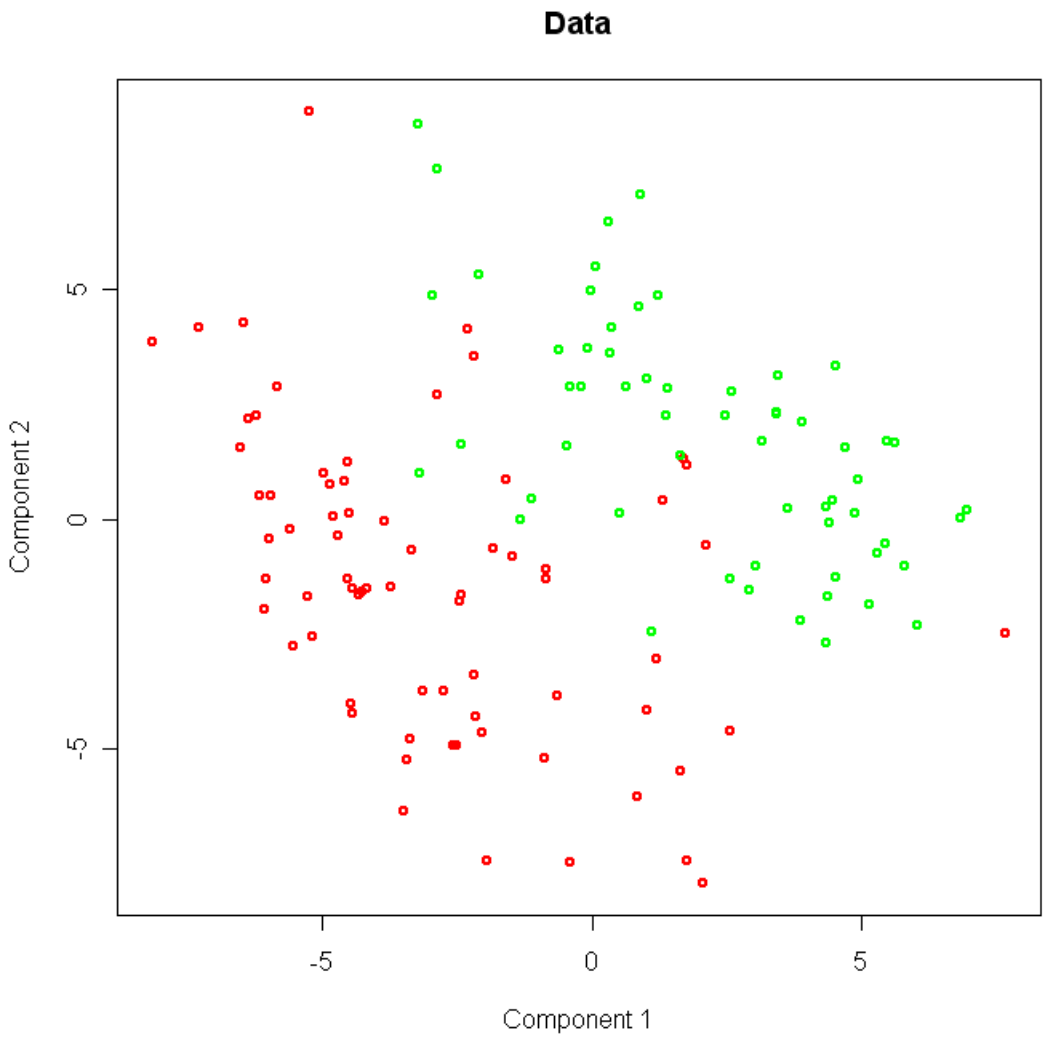
$$\widehat{sdg} = arg \min_{g \in S} E_{nn}(X_{test}; V)$$

Definition

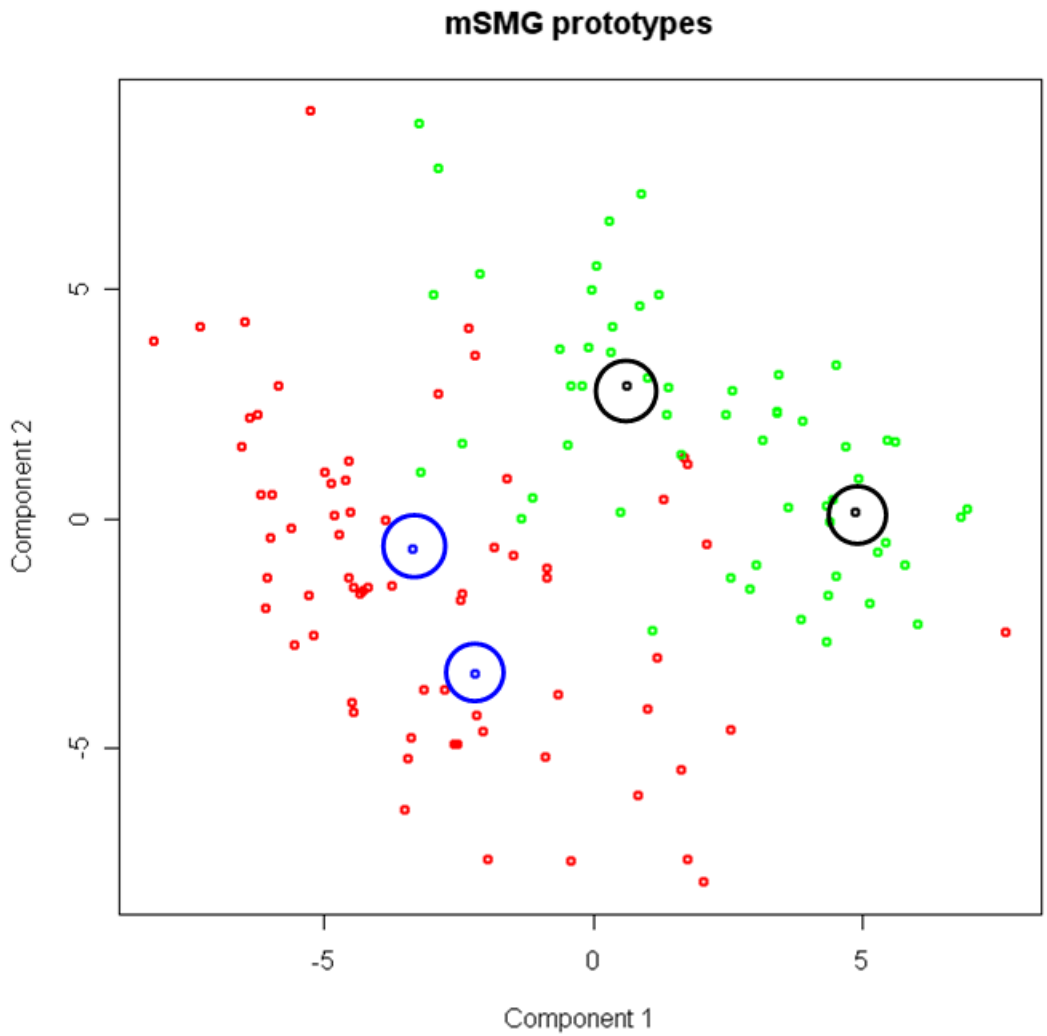
Generalized Discriminative Graphs (gdg)

$$\overline{gdg} = arg \min_{g \in U} E_{nn}(X_{test}; V)$$

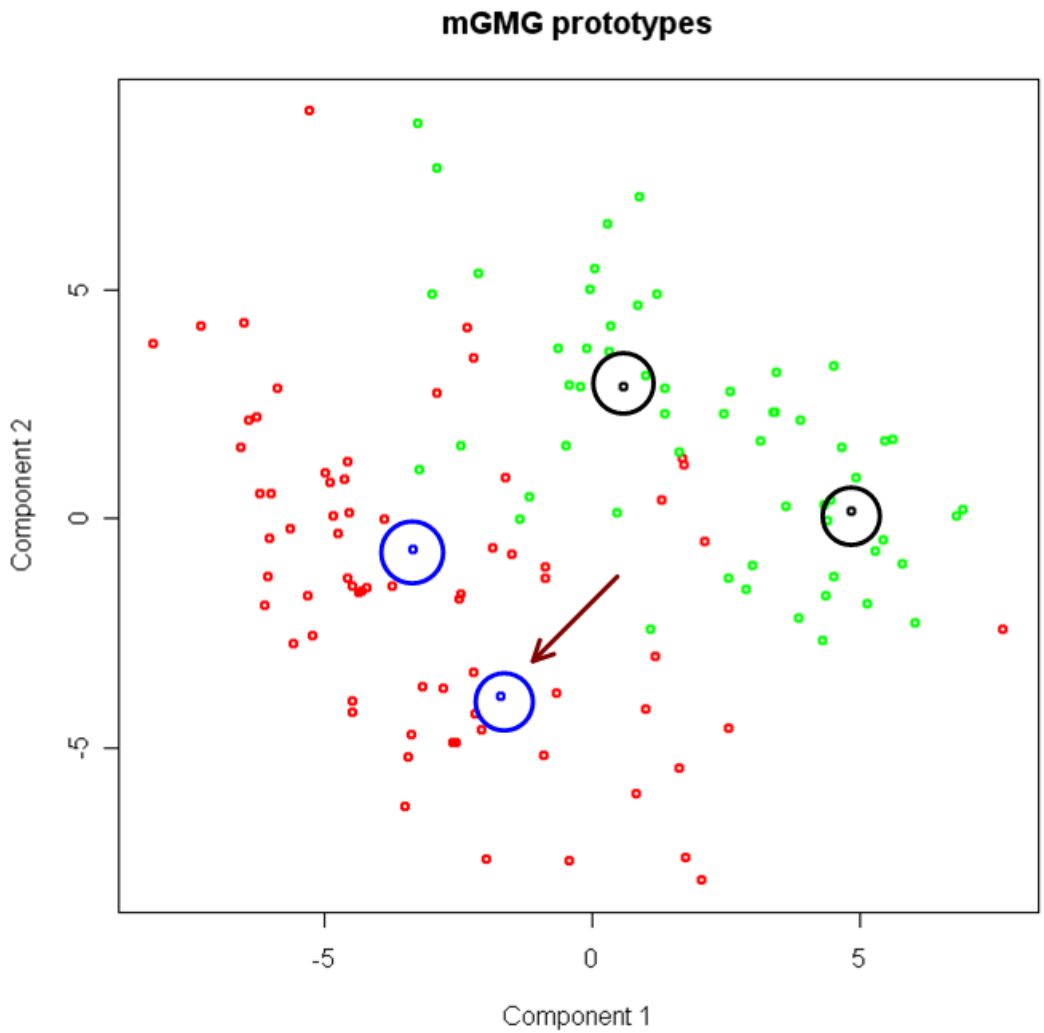
2 Classes (Red, Green): Dissimilarity space.



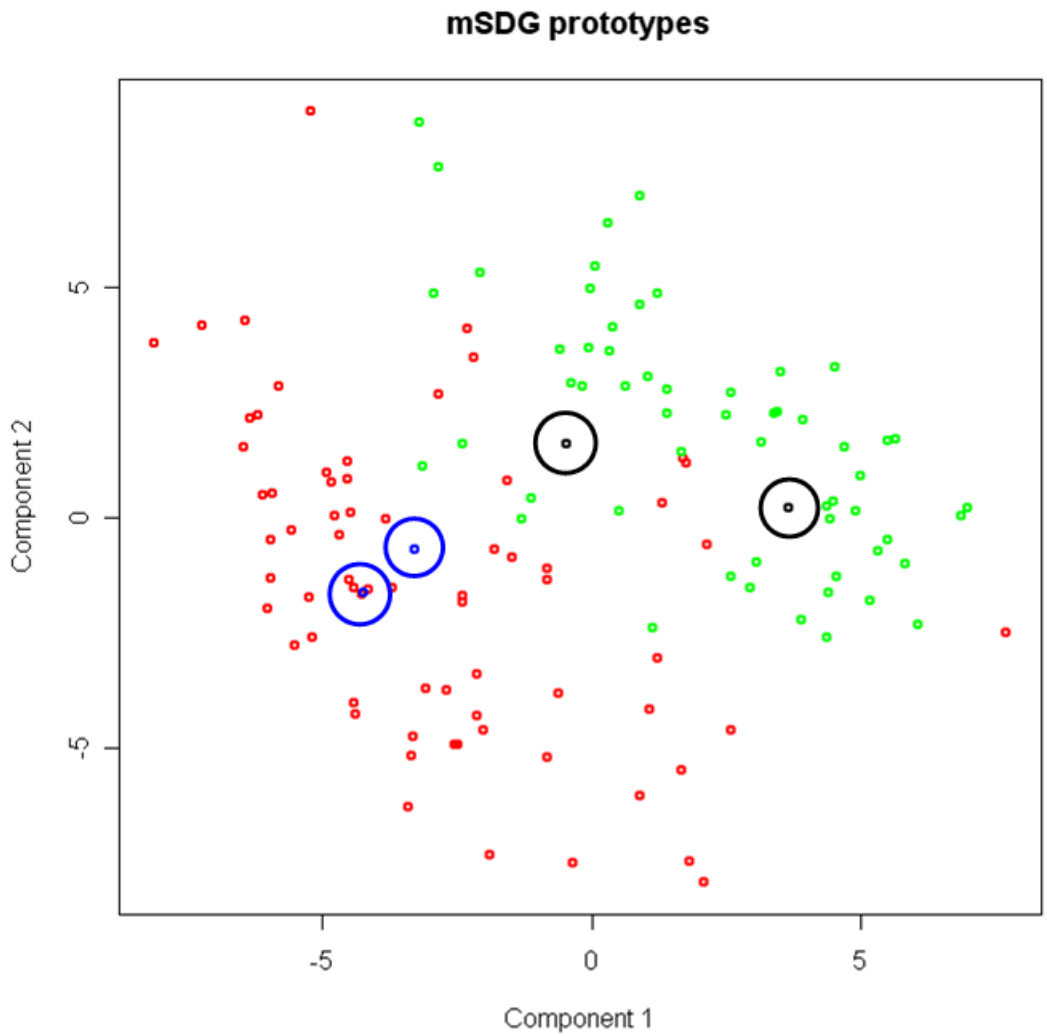
2 Classes : Set Median Graphs (SMG)



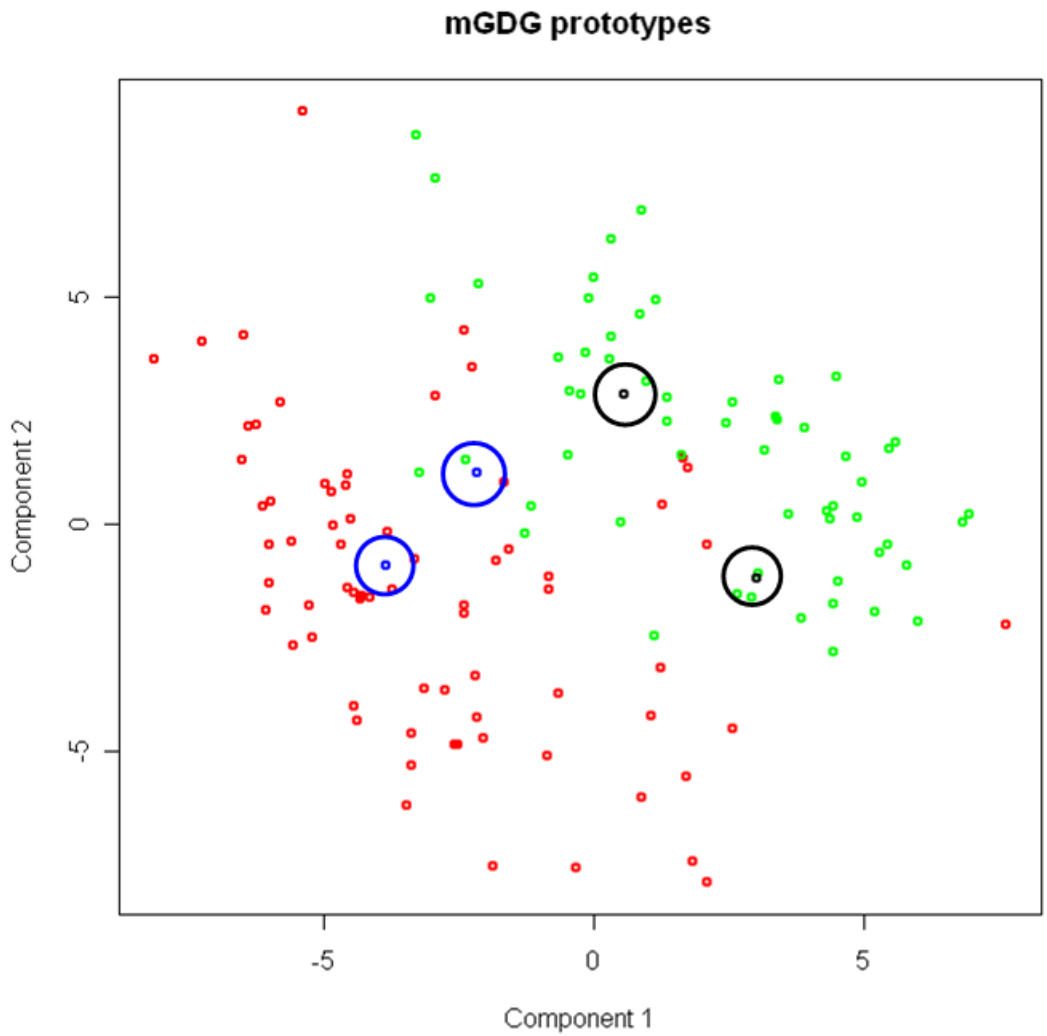
2 Classes : Generalized Median Graphs (GMG)



2 Classes : Set Discriminative Graphs (SDG)



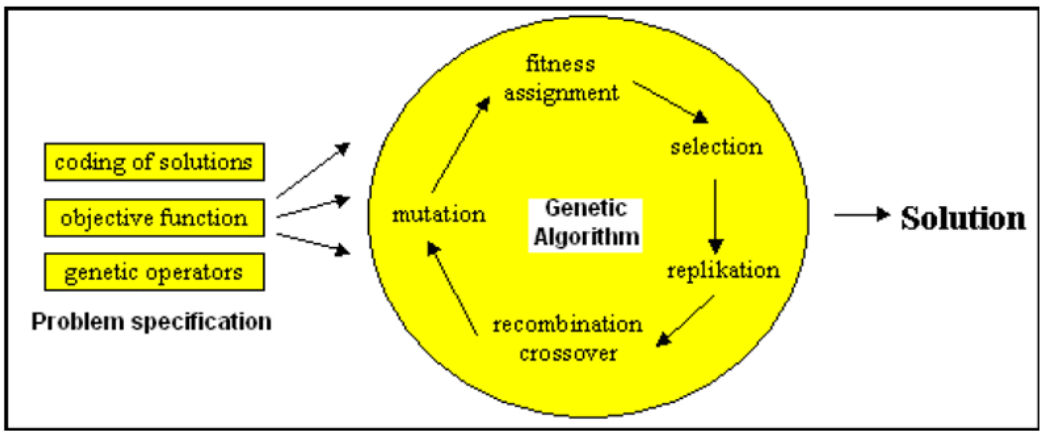
2 Classes : Generalized Discriminative Graphs (GDG)



Genetic Algorithm(GA) dedicated to graph prototypes

An optimization problem

- A criterion minimization (SOD or E_{nn});
- The M-prototypes searched is a NP-Complete problem [JIANG 01].
- GAs are well suited to solve this kind of non-analytical dilemma [GOLDBERG 89].



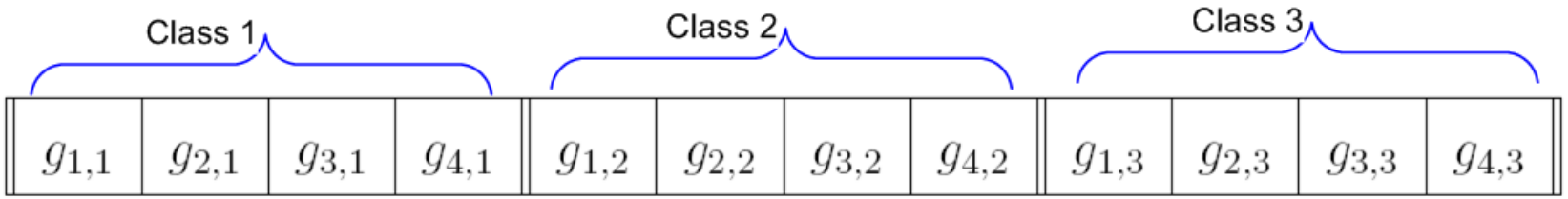
Generation of graph prototypes by GA

A framework for the generation of graph prototypes

- A specific encoding have to be done to take into account structured data.
- GA' basic functions are redefined : evaluation, mutation et cross-over.

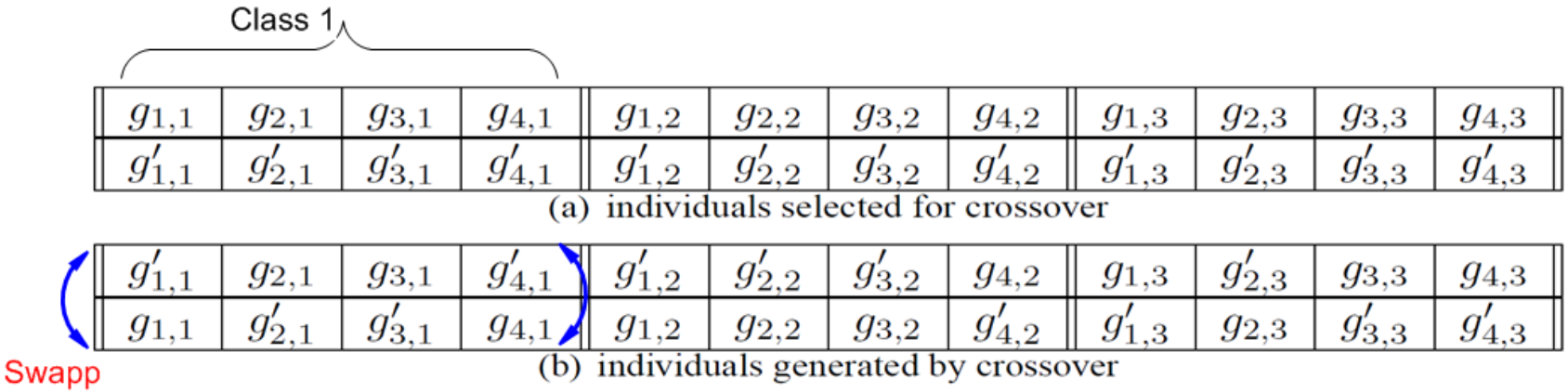
Individual encoding

- A vector V containing $M \times N$ graphs.



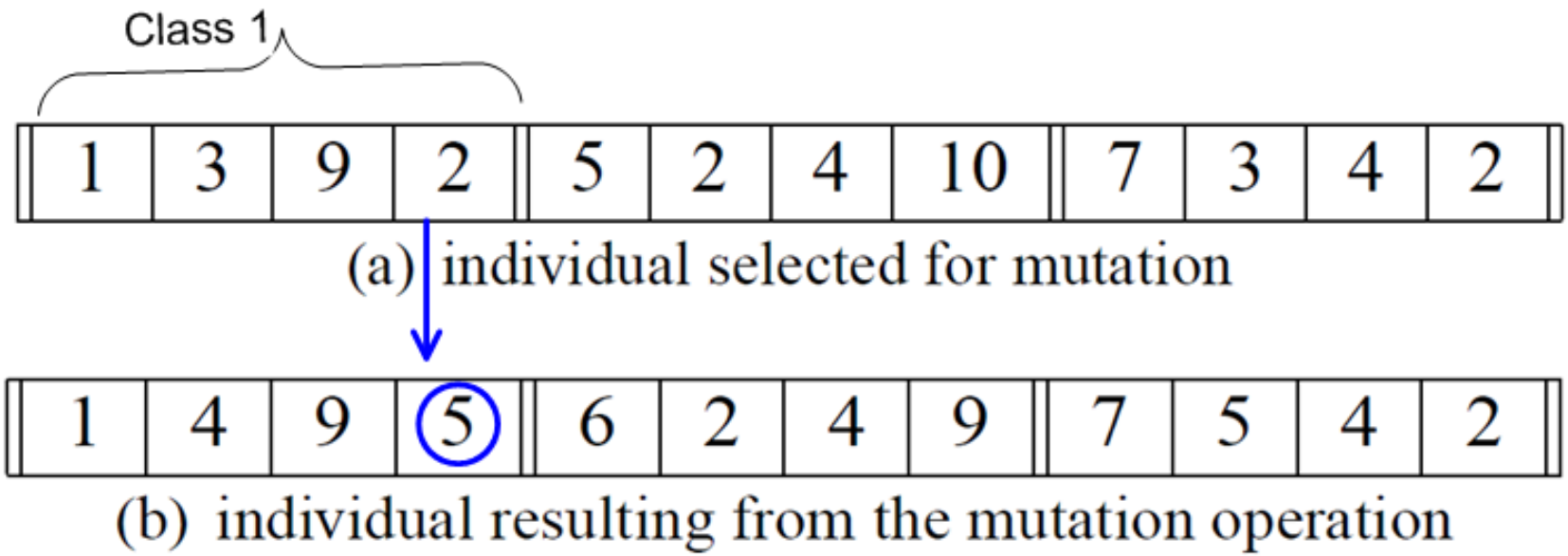
The cross-over function

- Exchange of set of graphs.



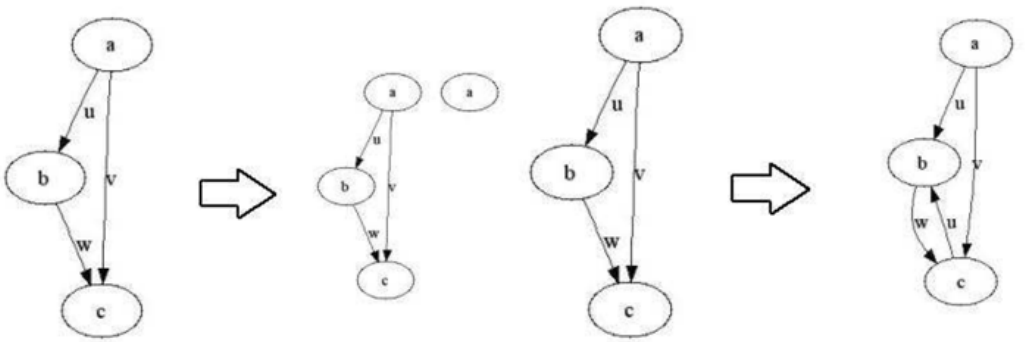
Mutation function for set graphs

- Set graphs selection $\in S$.
- A graph is randomly chosen from the training set.
- Picking up the graph number 5 of the class 1.

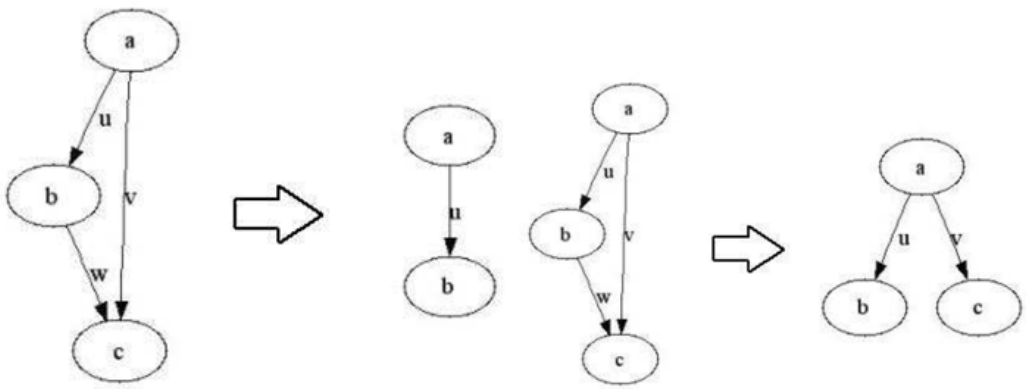


Mutation function for generalized graphs

- Graph generation $\in U$.



(a) Add node (b) Add edge



(c) Delete node (d) Delete edge

Fitness function

- Median Graphs : SOD minimization.
- Generalized Graphs : The error on a test base: $E_{nn}(X_{test}; V)$.

Both criteria lie on a dissimilarity measure between graphs:

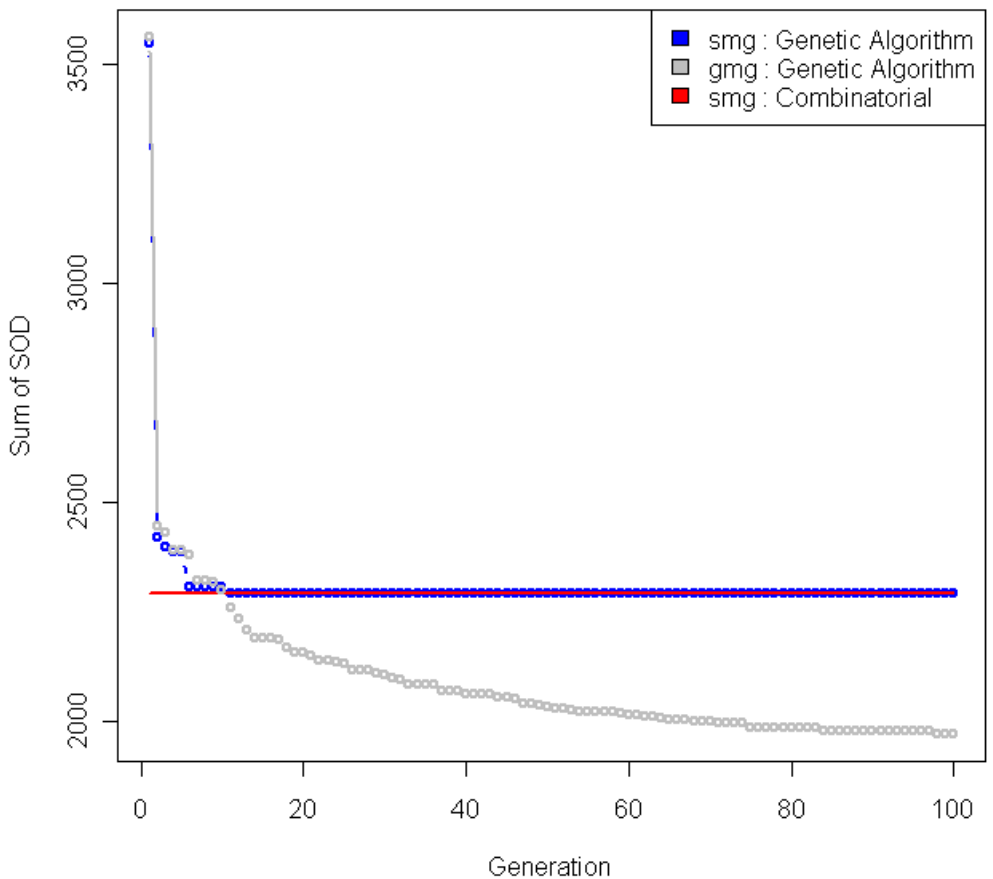
- The choice of a graph distance: $d(g_1, g_2)$
- Our choice: Graph Probing defined by Lopresti et Wilfong [LOP 03].
  - Fast, a $O(N)$ complexity in function of the number of nodes of the largest graphs.
  - A theoretical lower bound does exit with the graph edit distance [Bunke 98].
  - $gpd(g1, g2) <= 4.ed(g1, g2)$

Experimental results: The data sets

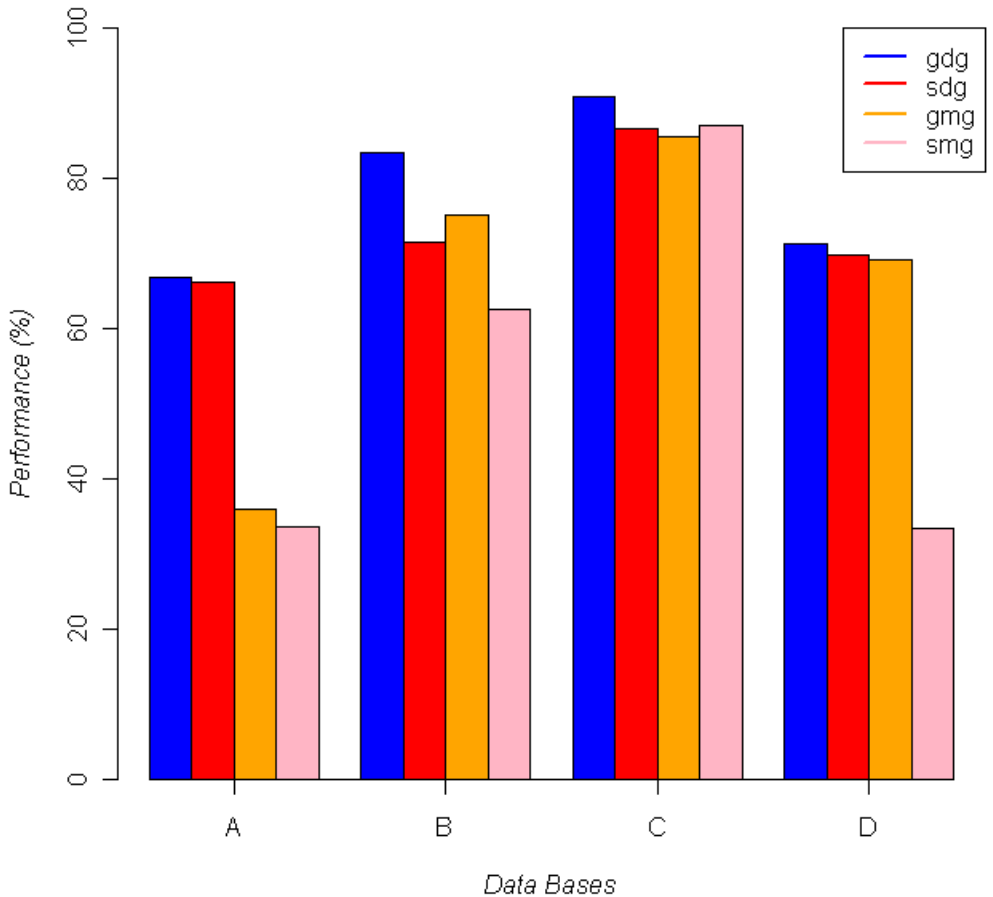
Table: Characteristics and specifications

	Synthetic (A)	GREC (B)	Ferrer (C)	Letter (D)
Number of classes (N)	50	10	32	15
$\mid$ *Training* $\mid$	10596	86	7200	3796
$\mid$ *Test* $\mid$	3532	56	2400	1266
$\mid$ *Validation* $\mid$	14101	56	3200	1688
Average number of nodes	12.03	5.56	8.84	4.7
Average number of edges	9.86	11.71	10.15	3.6
Average degree of nodes	1.63	4.21	1.15	1.3

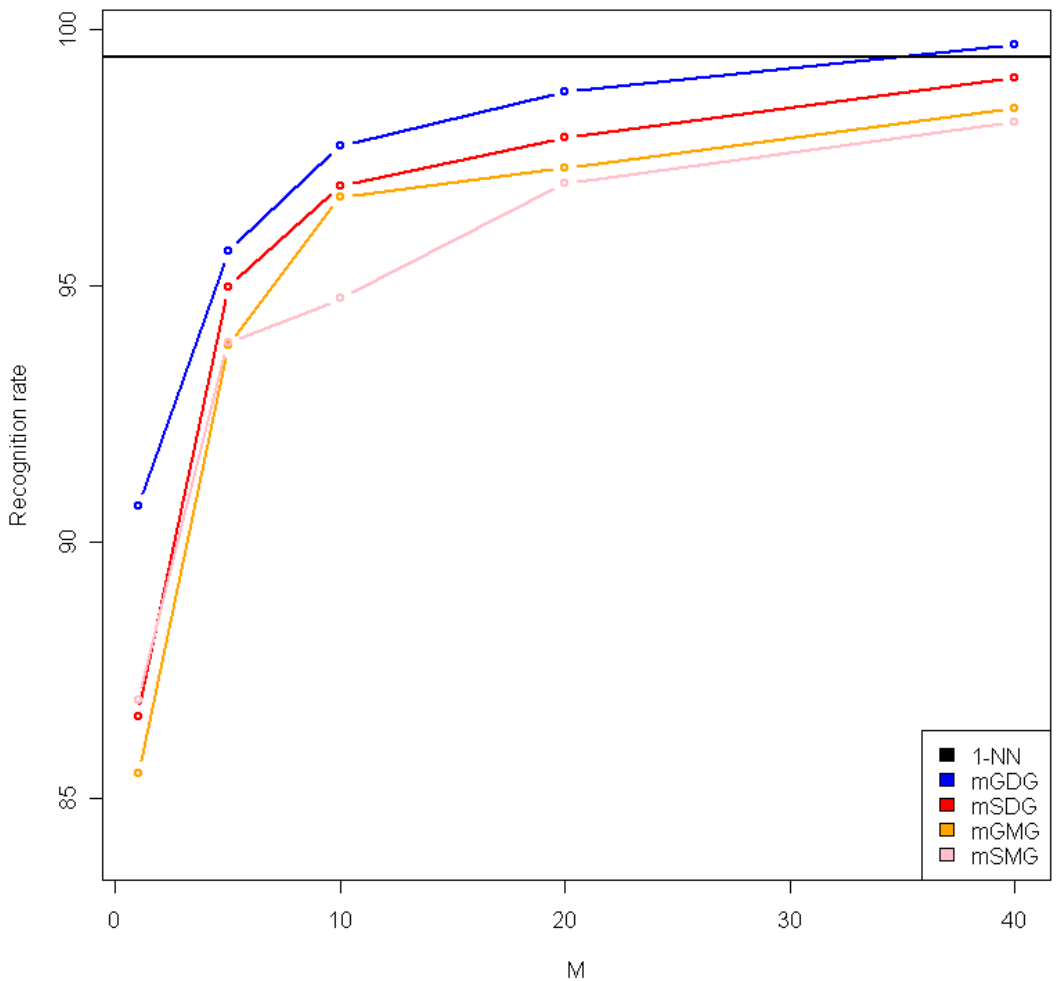
Convergence of the Genetic Algorithm



Classification using one prototype per class



Classification using M prototypes per class



Conclusion

Contribution

- A genetic algorithm dedicated to graph prototypes. Generation of different kind of prototype.
- The possibility to produce M prototypes per classes.

Our conclusions

- Generalized Median Graph has a better modeling ability than Set median Graph.
- Generating M prototypes per class increase the performance in classification.
- Generalized Discriminative Graphs: $E_{nn}(X_{test}; V) <= E_{nn}(X_{test}; X_{tr})$

Food for thoughts

- To introduce probabilistic models into the GA' functions.
- To specialize GA' function for the symbol application.

- Thanks for your attention ;-)