Possibilistic Clustering for Data Mining in High Energy Physics

F. Masulli(1,3) A.M. Massone(1,2) L. Studer(2)

(1) INFM Istituto Nazionale per la Fisica della Materia -
Via Dodecaneso 33, 16146 Genova, Italy
(2) IPHE Université de Lausanne
BSP - 1015 Lausanne, Switzerland
(3) DISI Dipartimento di Informatica e di Scienze dell'Informazione Università di Genova - Via Dodecaneso 35, 16146 Genova
masulli@disi.unige.it

Introduction

- In many applications we deal with mountains of unstructured data, and we want to mine from them salient information useful for our needs.
- Clustering algorithms can make possible to filter unstructured data bases by using a-priori knowledge expressed in form of models
- Often, the probabilistic constraint used by some popular clustering algorithms gives rise to an unpleasant interference among the models.

In this paper we enlighten the good characteristics of the Possibilistic Approach to Clustering, i.e.:

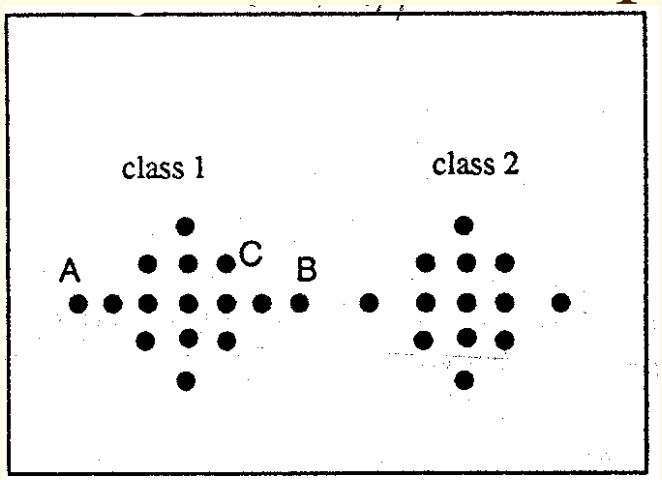
- small interference among the models;
- high robustness to outliers

Anomalies due to the probabilistic constraint

$$\sum_{j=1}^{c} u_{jk} = 1$$

where u_{jk} is the membership of point k-th to cluster j-th.

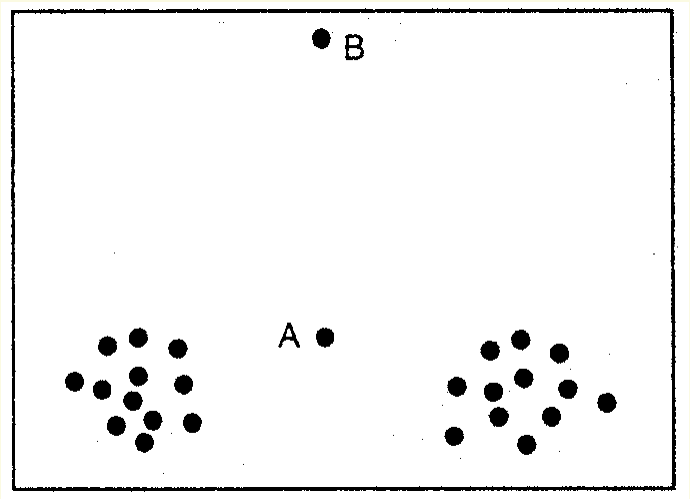
- Anomalies -Example1



a,b symmetric

$$u_{c_1}(a) \neq u_{c_1}(b)$$

✔ Anomalies - Example 2



a,b outliers

$$u_{c_1}(a) = u_{c_1}(b) = u_{c_2}(a) = u_{c_2}(b) = .5$$

but b is less representative than a

Possibilistic C-Means Algorithm

(Keller and Krishnapuram,1993,1996)

- membership function: **degree of typicality**
- minimization of the HCM functional: relax the probabilistic constraint to a possibilistic constraint:

$$u_{jk} \in [0,1] \quad \forall\, j,k;$$

$$0 < \sum_{k=1}^{n} u_{jk} < n \quad \forall\, j;$$

$$\vee_{j} u_{jk} > 0 \quad \forall\, j.$$

PCM-II cost function (KK, 1996)

$$J(U,Y)=\sum_{j=1}^{c}\sum_{k=1}^{n}u_{jk}E_j(x_k)+\sum_{j=1}^{c}\eta_j\sum_{k=1}^{n}(u_{jk}\log u_{jk}-u_{jk})$$

- the first term is the HCM cost function

  $E_j(x_k)$ is the square of the Euclidean distance

- the second term is a regularization term

  η_j is the **intracluster** distance, i.e. a generalization of the cluster's variance. It is the zone of influence of a model.

$J(U,Y)$ can be interpreted also as a generalized free entropy

PCM-II learning rules (KK,1996)

$J(U,Y)$ is optimized by:

$$u_{jk} = \exp\left\{-\frac{E_j(x_k)}{\eta_j}\right\} \qquad y_j = \frac{\sum_{k=1}^{n} u_{jk} x_k}{\sum_{k=1}^{n} u_{jk}}$$

Those two equations can be interpreted as learning rules for a learning by epoch procedure.

The PCM-II works as a refining algorithm and must be initialized by another clustering algorithm.

The LHCb experiment at CERN

In 2005, the *Large Hadron Collider* (LHC) at CERN Geneva will be commissioned. Proton against proton will collide at an energy of 14 TeV.

One of the four experiments, LHCb is dedicated to the study of CP violation in the B meson system, an ingredient for the explanation of the matter dominance in our Universe.

LHCb will use various detectors; among them two *Ring Imaging Cherenkov* (RICH) counters which will identify the type of stable charged particles (π, K, p, μ, and e) produced in the decay of B mesons.

Cherenkov angle

If a charged particle goes through a dielectric material at a speed greater then the speed of light in this material, *photons are emitted at a characteristic angle* θ *from the charged particle flypath*.

$$\cos\theta = \frac{1}{\beta\, n}$$ Cherenkov angle

- β is the ratio of the speed of the particle to the speed of light in vacuum
- n is the index of refraction of the medium.

Rings in RICH counters

By a clever arrangement of focusing mirrors, it is possible to collect and detect *the Cherenkov light emitted by the charged particle on a surface* where this light *forms circular rings*.

The ring diameter is a function of θ *and then of* β.
Assuming that another LHCb detector measures the momentum of the *charged particle*, its *type* can be derived.

Pattern Recognition problem

- Rings with possible elliptical shape
- Not complete rings
- Outliers - noise
- High data flow

The ring pattern recognition has to be robust and very tolerant to imprecision.

Data flow

- One collision every 25 ns $\Rightarrow$ few tens of charged particles, i.e. rings
- Each ring about 25 photons
- Purity $\Leftrightarrow$Efficiency
- Data flow entering Level in is 4GB/s
- Final Data storage rate is about 20Mb/s

Trigger Levels

- *Level 0* trigger 40 MHz - selects about 1/40 events - latency 3.2 micro s
- *Level 1* trigger 1 MHz - selects about 1/25 events - latency 120 micro s
- *Level 2* trigger 40kHz - selects about 1/8 events - latency 10ms.
- *Level 3* trigger 200 Hz - selects about 1/25 events, latency 200 ms

Level 3 is a complete analysis. RICH rings are examined only at this level 3 .

The ring detection algorithm should be implemented at Level 2.

Possibilistic C-Spherical Shell algorithm (Krishnapuram et al 1995)

- prototypes (c_j, r_j)

  c_j is the center of a ring

  r_j is the radius of a ring

- distance of a point to a ring prototype

$$E_j(x_k) = d_{jk}^2 = d^2(x_k; c_j, r_j) = (\|x_k - c_j\| - r_j)^2$$

- objective function

$$J(U, L) = \sum_{j=1}^{c} \sum_{k=1}^{n} u_{jk} E_j(x_k) + \sum_{j=1}^{c} \eta_j \sum_{k=1}^{n} (u_{jk} \log u_{jk} - u_{jk})$$

Note:

The distance of a point to a prototype can be rewritten as:

$$d_{jk}^2 = p_j^T M_k p_j + v_k^T p_j + b_k$$

where:

$$b_k = \left(x_k^T x_k\right)^2 \qquad v_k = 2\left(x_k^T x_k\right) y_k \qquad y_k = \begin{bmatrix} x_k \\ 1 \end{bmatrix}$$

$$M_k = y_k y_k^T \qquad p_j = \begin{bmatrix} -2c_j \\ c_j^T c_j - r_j^2 \end{bmatrix}$$

Possibilistic C-Spherical Algorithm: learning rules

The PCSS objective function is minimized by

$$p_j = -\frac{1}{2}\left(H_j\right)^{-1}\omega_j$$

where $$H_j = \sum_{k=1}^{n} u_{jk} M_k \qquad \omega_j = \sum_{k=1}^{n} u_{jk} v_k$$

and by

$$u_{jk} = \exp\left\{-\frac{E_j(x_k)}{\eta_j}\right\}$$

Basic Possibilistic C-Spherical Shell Algorithm

- Initialization

REPEAT

- Calculate $H_j \quad \omega_j \quad \forall j$
- Compute $p_j \quad \forall j$
- Calculate u_{jk}

UNTIL *stopcond* = *T*

PCSS Initialization

♣ Fuzzy C-Means Algorithm (Bezdek, 1982) $\Rightarrow c_j$

For each cluster $\Rightarrow$ histogram on X axis $\Rightarrow$ estimation of the diameter of the ring

Let be $r_j \pm pr_j$ the distance from the prototype ring at which $u_{jk} = .5$

$$\Rightarrow \quad \frac{1}{2} = \exp\left\{-\frac{E_j(x_k)}{\eta_j}\right\} = \exp\left\{-\frac{p^2 r_j^2}{\eta_j}\right\}$$

then

♣ $$\eta_j = \frac{p^2}{\ln 2} r_j = 1.44 \cdot p^2 r_j$$

e.g. $\eta_j = .13 \cdot r_j$ if $p = .3$

PCSS *stopcond*

PCSS terminates if

the changes in the prototypes are less than an assigned threshold

Iterative Possibilistic C-Spherical Shell Algorithm (IPCSS)

- $c = 3 \cdot \hat{c}$

REPEAT

- PCSS (c)
- Filter clusters with small fuzzy cardinality
- Collapse clusters with overlapping centroids
- Generate a new database filtering points classified as member of a cluster or as outlier
- $c = .5 \cdot c$

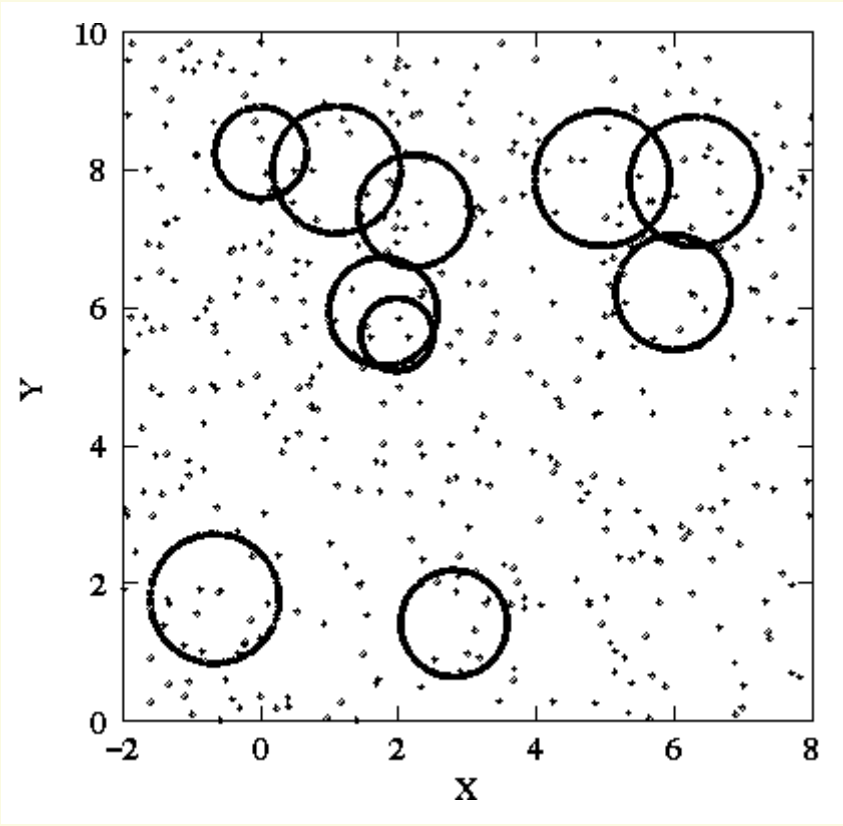
UNTIL (0 centroids are found)

Data Base

10 rings

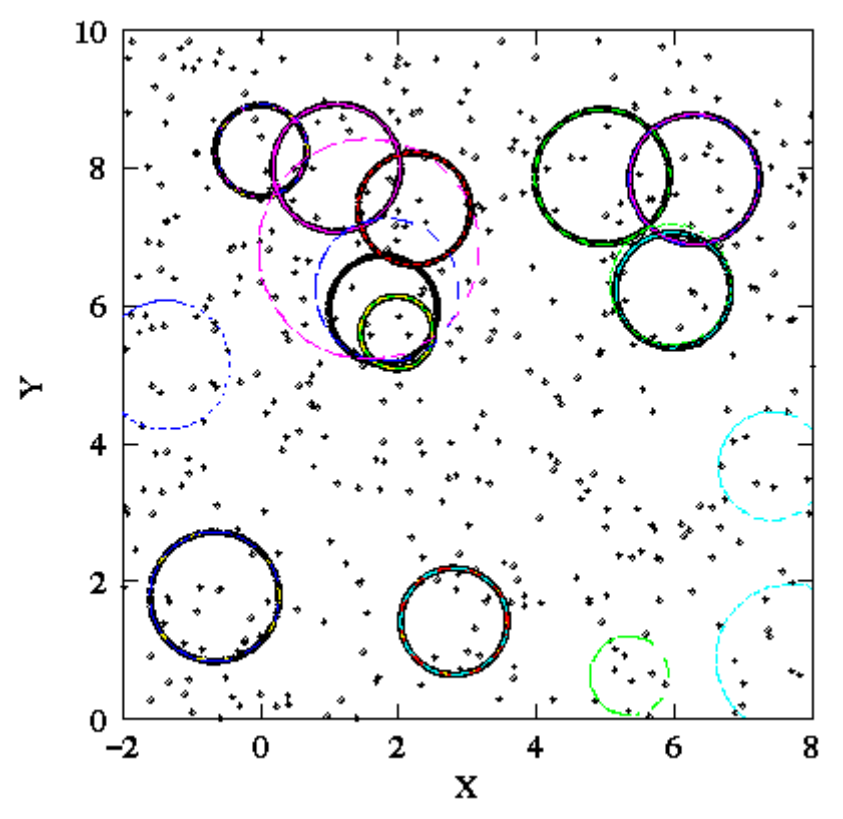
- 500 points each
- radii .5-1

noise 10%



Iteration 1 - Step 1

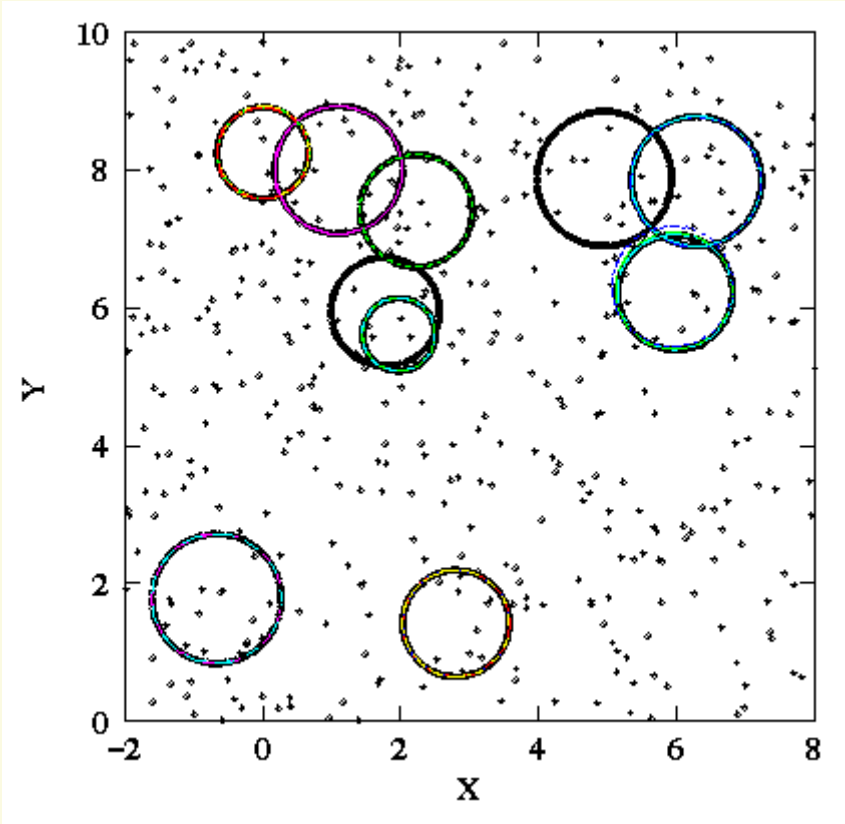
- ✔ PCSS solution using 30 prototypes



Iteration 1 - Step 2

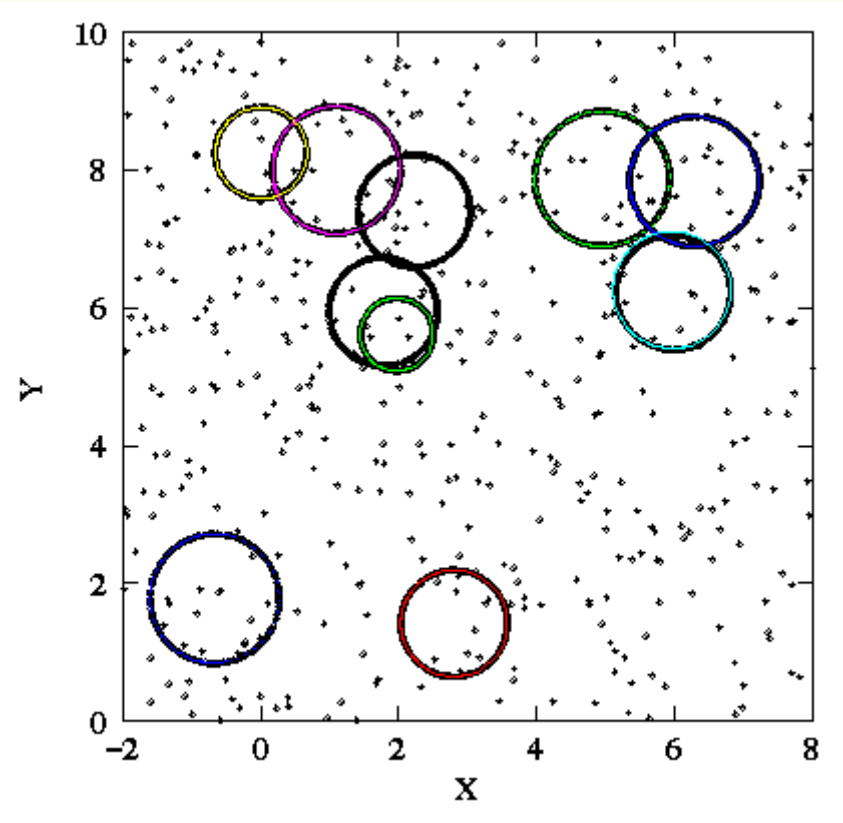
- 24 filtered prototypes using cardinality threshold

$$\tau_0 = 400$$



Iteration 1 - Step 3

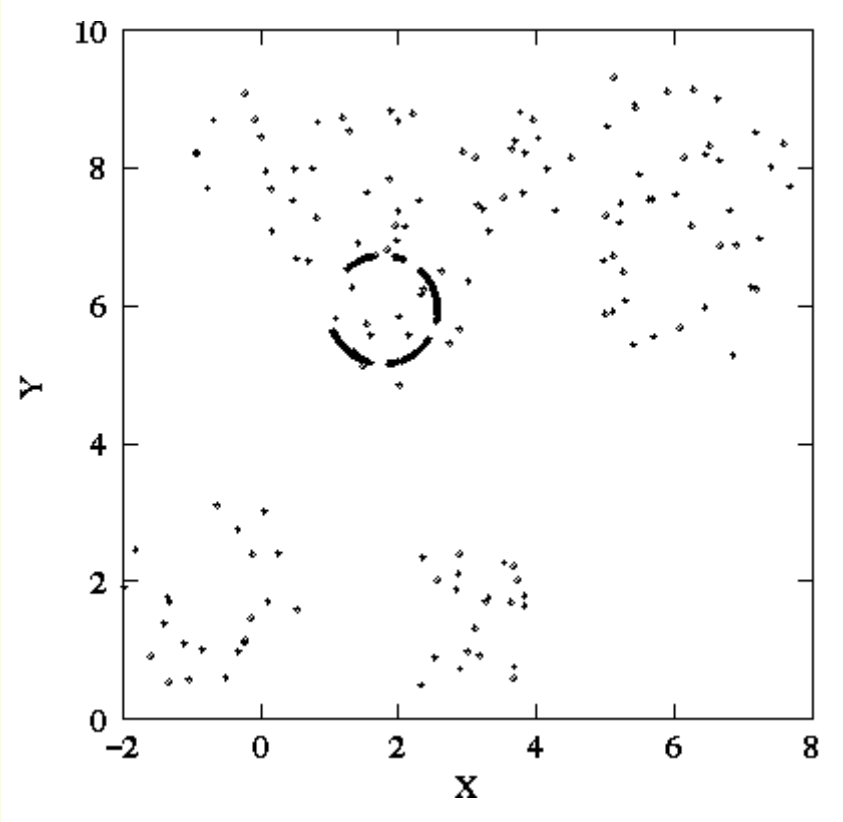
- ✔ Collapse ⇒ 9 rings



Iteration 1 - Step 4

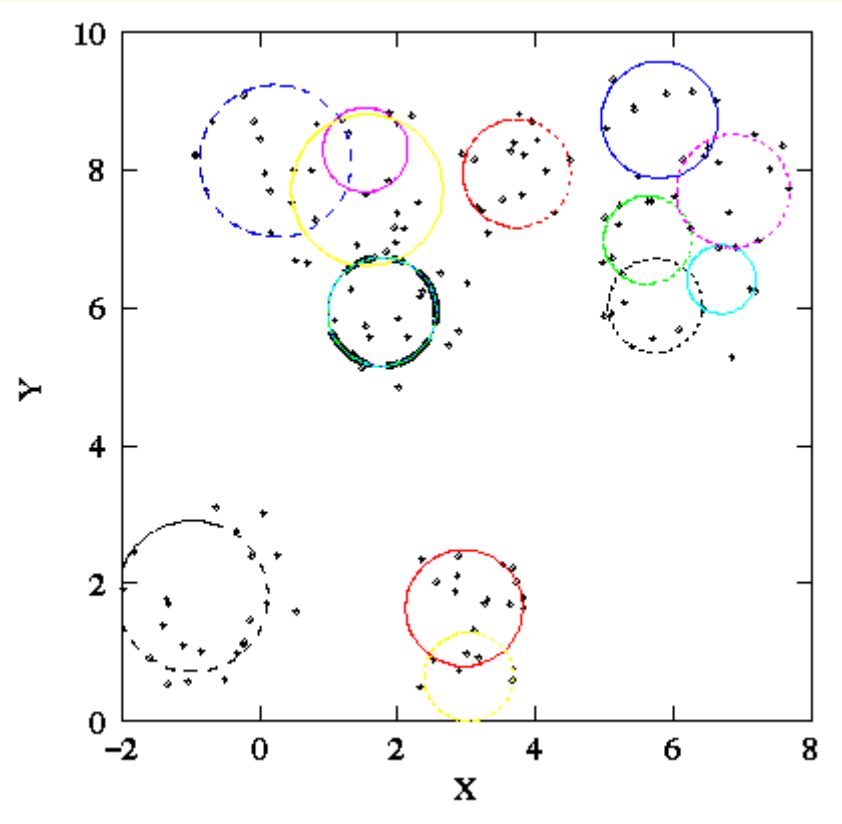
- Cancel bad/well classified points

$$\tau_1 = .96 \qquad \tau_2 = .1$$



Iteration 2 - Step 1

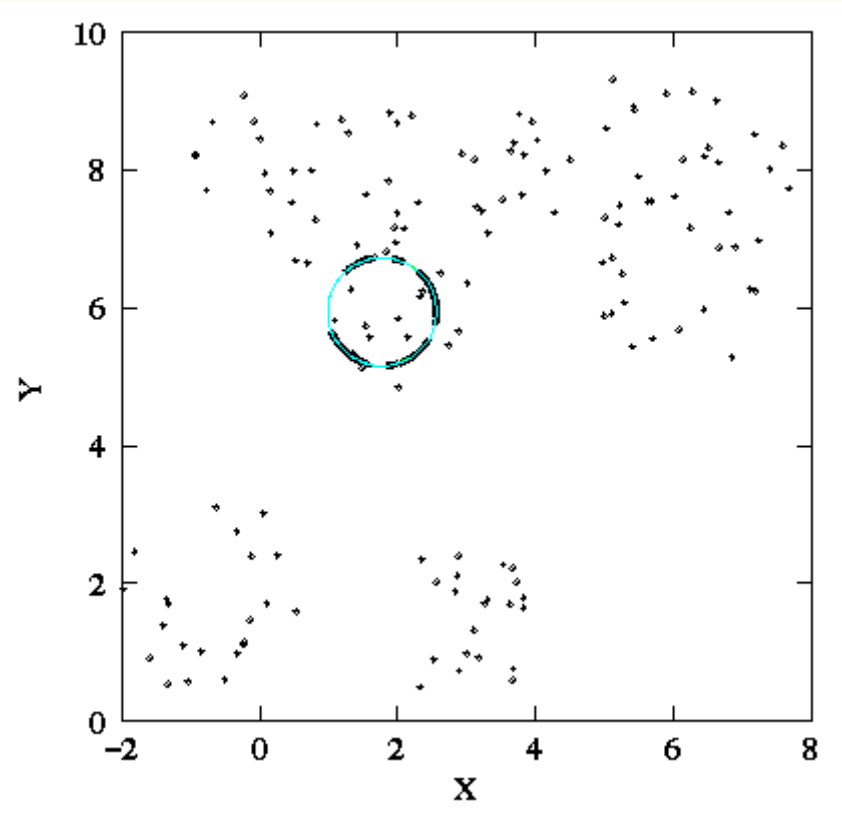
- PCSS solution using 15 prototypes



Iteration 2 - Step 2

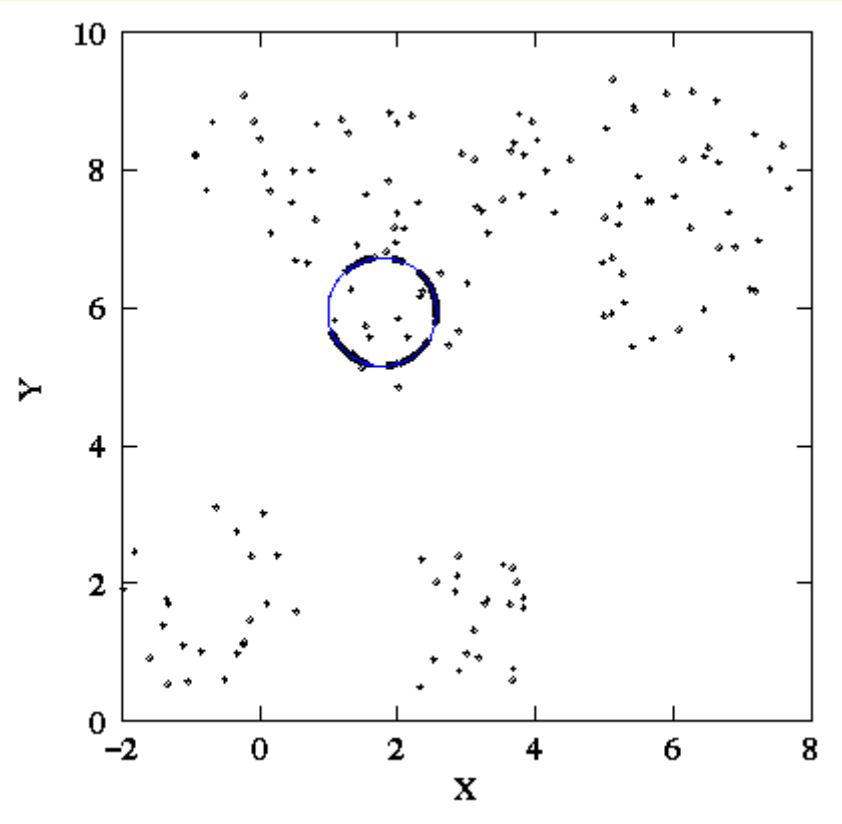
- 6 filtered prototypes using cardinality threshold

$$\tau_0 = 300$$



Iteration 2 - Step 3

- ✔ Collapse ⇒ 1 ring



IPCSS solution