Sensor/Marker Selection for Diagnosis based on a Fuzzy Feature Selection Approach

L.HEDJAZI , M-V LE LANN, T. KEMPOWSKY,
J. AGUILAR

LAAS-CNRS ,Toulouse

S3 ENSAM March 25, 2011

Contents

I. Introduction 1

- **classification or pattern recognition techniques offer the advantages to be data-driven approach**

- **These techniques have been shown very effective for medical diagnosis as cancer prognosis characterized by:**
  - High dimensionality of data (e.g. thousands of genes)
  - Data heterogeneity (e.g. Qualitative, quantitative, interval)

- **These techniques have also been shown effective for fault detection and diagnosis of complex industrial processes.**

- **Despite their behavioral difference, both domains (industrial processes and medical diagnosis) exhibit many common practices:**

→ problem of marker selection for medical diagnosis

→ problem of sensor selection for industrial process diagnosis

I. Introduction 2

- **A novel methodology enables to handle simultaneously both problems regardless of their own characteristics:**

→ Copes with the problem of high dimensionality based on classical optimization methods.

→ Handles appropriately heterogeneous data (quantitative, qualitative, interval) .

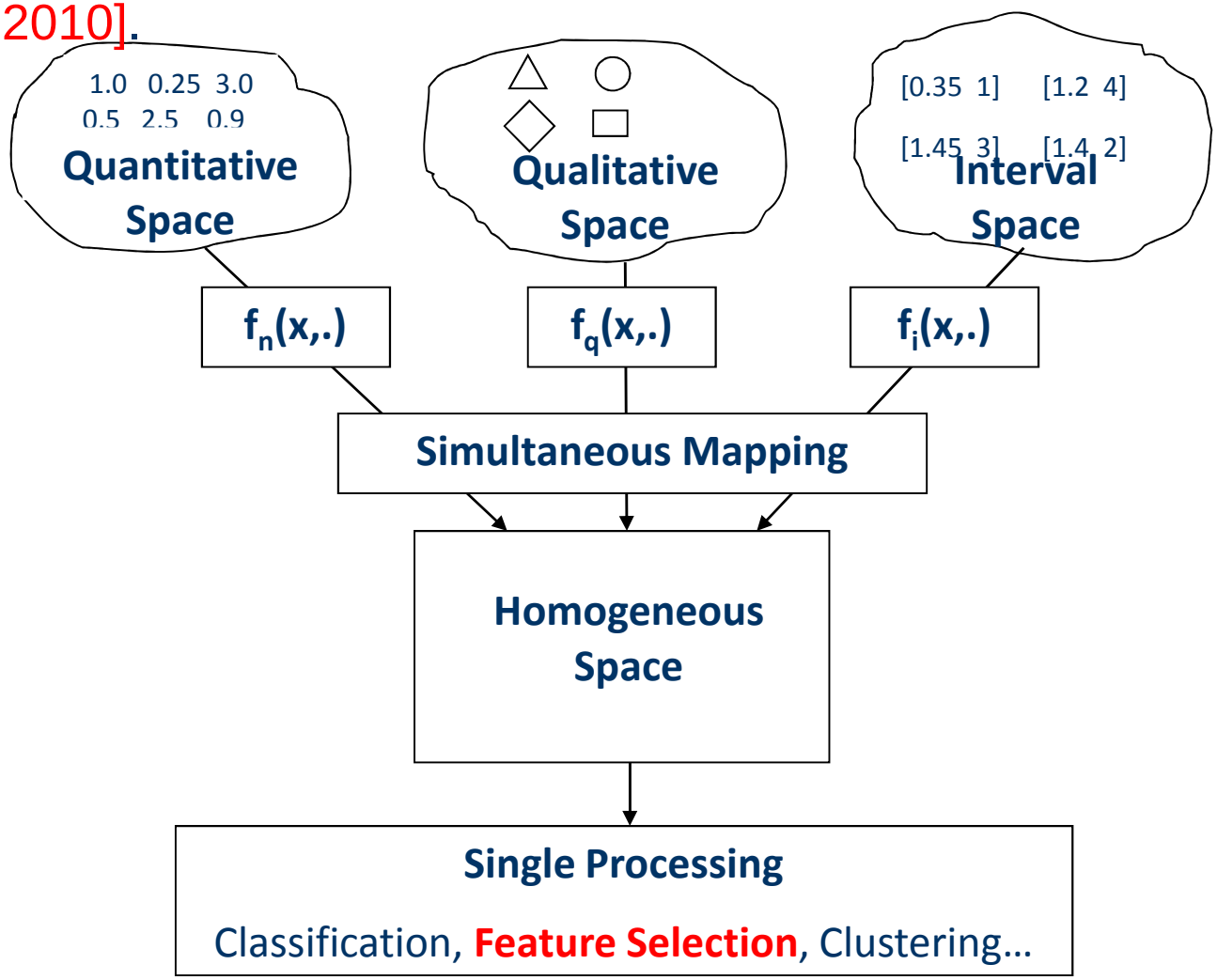
Relevant example: interval representation of data can improve classification processing of clinical data in medical diagnosis as well as to process noisy or uncertain industrial measurements.

- Application to derivation of hybrid markers for cancer prognosis.

- Application to sensor selection for fault diagnosis of pharmaceutical synthesis process in a new intensified heat-exchanger reactor.

II. Fuzzy Feature Selection

- **MEMBAS (MEmbership Margin Based feAture Selection):** [Hedjazi et al., 2010].



II.1 SM:Simultaneous Mapping

- Membas enables to process three types of features: **Quantitative, Qualitative** and **Interval .**
- In Fuzzy Logic framework, Simultaneous mapping can be performed through a Feature Fuzzification according to each feature type.
- *Based on a data-driven procedure,* ℓ fuzzy partitions $\{ mff_1^i, \ldots, mff_\ell^i \}$ are obtained for the i^{th} feature to each existing class k :

$$mff_k^i = \mu_k^i (x_i, \theta_{ki})$$

- **Quantitative type features (**similarity semantic):

$$\mu_k^i \left[x_i \middle| \rho_k^i, \varphi_k^i \right] = {\varphi_k^i}^{1-\left|x_i-\rho_k^i\right|} \left(1-\varphi_k^i\right)^{\left|x_i-\rho_k^i\right|} \quad \text{where} \quad \varphi_k^i = \frac{1}{m_k} \sum_{j=1}^{j=m_k} x_i^j$$

II.1 SM:Simultaneous Mapping

Interval type features (similarity semantic):

$$S(A,B)=\frac{1}{2}\left(\frac{\varpi[A\cap B]}{\varpi[A\cup B]}+1-\frac{\partial[A,B]}{\varpi[U]}\right)$$

where $\partial[A,B]=\max\left[0,\left(\max\{a^-,b^-\}-\min\{a^+,b^+\}\right)\right]$ *and* $\varpi[X]=upper.bound(X)-lower.bound(X)$

Therefore,

$$\mu_k^i(x_i)=S(x_i,\rho_k^i)$$

where

$$\rho_k^{i-}=\frac{1}{m_k}\sum_{j=1}^{m_k}x_i^{j-} \quad \textit{and} \quad \rho_k^{i+}=\frac{1}{m_k}\sum_{j=1}^{mk}x_i^{j+}$$

↗ **Qualitative type features** (Uncertainty semantic):

$$\mu_k^i(\chi_i)=\left(\Phi_{k1}^i\right)^{q_{i1}}*\cdots*\left(\Phi_{kMi}^i\right)^{q_{iMi}}$$

Where Φ_{kM}^i is the frequency of the Mth modality in the class C_k

and

$$q_j^i=\begin{cases}1 & if \quad x_i=Q_j^i\\ 0 & if \quad x_i\neq Q_j^i\end{cases}$$

II.1 SM: Simultaneous Mapping

Let $D = \{x_n, C_k\}_{n=1}^{N} \in X \times C$ be a dataset, where N is the number of patterns (items) and $x_n=[x_{n1},x_{n2},\ldots,x_{nm}]$ is the n^{th} pattern.

- A natural result of the previous fuzzification step is a common membership space for heterogeneous features, i.e. a Membership Degree Vector (MDV) of pattern to each class:

$$U_{nc_k} = \left[\mu_k^1(x_{n1}),\ \mu_k^2(x_{n2}),\ldots,\ \mu_k^m(x_{nm})\right]^T;\quad k=1,2,\ldots,l$$

where $\mu_k^i(x_{ni}) = \mu_k^i(x_i = x_{ni})$

- Each MDV can be considered as a discrete fuzzy subset.

II.2 SP: MEMBAS for binary class problems

- A membership margin is defined for each pattern $x_n \in c$:

$$\beta_n = \psi\left(U_{nc}\right) - \psi\left(U_{n\tilde{c}}\right)$$

- Where $\psi(U_{nc_k}) = \sum_i \mu_k^i(x_{ni})$ is the scalar cardinality of the fuzzy subsets described by MDVs.

- Pattern x_n is considered correctly classified if $\beta_n > 0$.

- Weighted adequacy assignment concept through the scalar cardinalities:

$$\Psi(U_{nc_k}/W_f) = W_f^T U_{nc_k} = \sum_i w_{fi}\, \mu_k^i(x_{ni})$$

- A weighted membership margin can be defined as:

$$\beta_n = \psi(U_{nc}/W_f) - \psi(U_{n\tilde{c}}/W_f)$$

II.2 SP: MEMBAS for binary class problems

- A margin-based objective function has been defined so that the averaged membership margin in the resulted weighted membership space is maximized:

$$\max_{w_f} \sum_{n=1}^{N} \beta_n(w_f) = \sum_{n=1}^{N} \{ \sum_{i=1}^{m} w_{fi} \mu_c^i(x_{ni}) - \sum_{i=1}^{m} w_{fi} \mu_{\tilde{c}}^i(x_{ni}) \}$$

$$\text{s.t. } : \| w_f \|_2^2 = 1 \text{ , and } \quad w_f \geq 0$$

- A closed-form solution using Lagrangian

$$w_f^* = \frac{s^+}{\| s^+ \|}$$

where $s = \sum_{n=1}^{N} \{ U_{nc} - U_{n\tilde{c}} \}$

with $s^+ = [\max(s_1, 0), \ldots, \max(s_m, 0)]^T$

II.2 SP: MEMBAS for multiclass problems

- Complex process are characterized by a big number of classes

→ An extension of Membas to multiclass problems is needed.

- A membership margin definition for multiclass problems was used:

$$\beta_n = \min_{\{\tilde{c} \in C, \tilde{c} \neq C(\mathrm{x}_n)\}} \left\{ \Psi\left(\mathrm{U}_{nc}\right) - \Psi\left(\mathrm{U}_{n\tilde{c}}\right) \right\}$$

- This leads to a similar solution

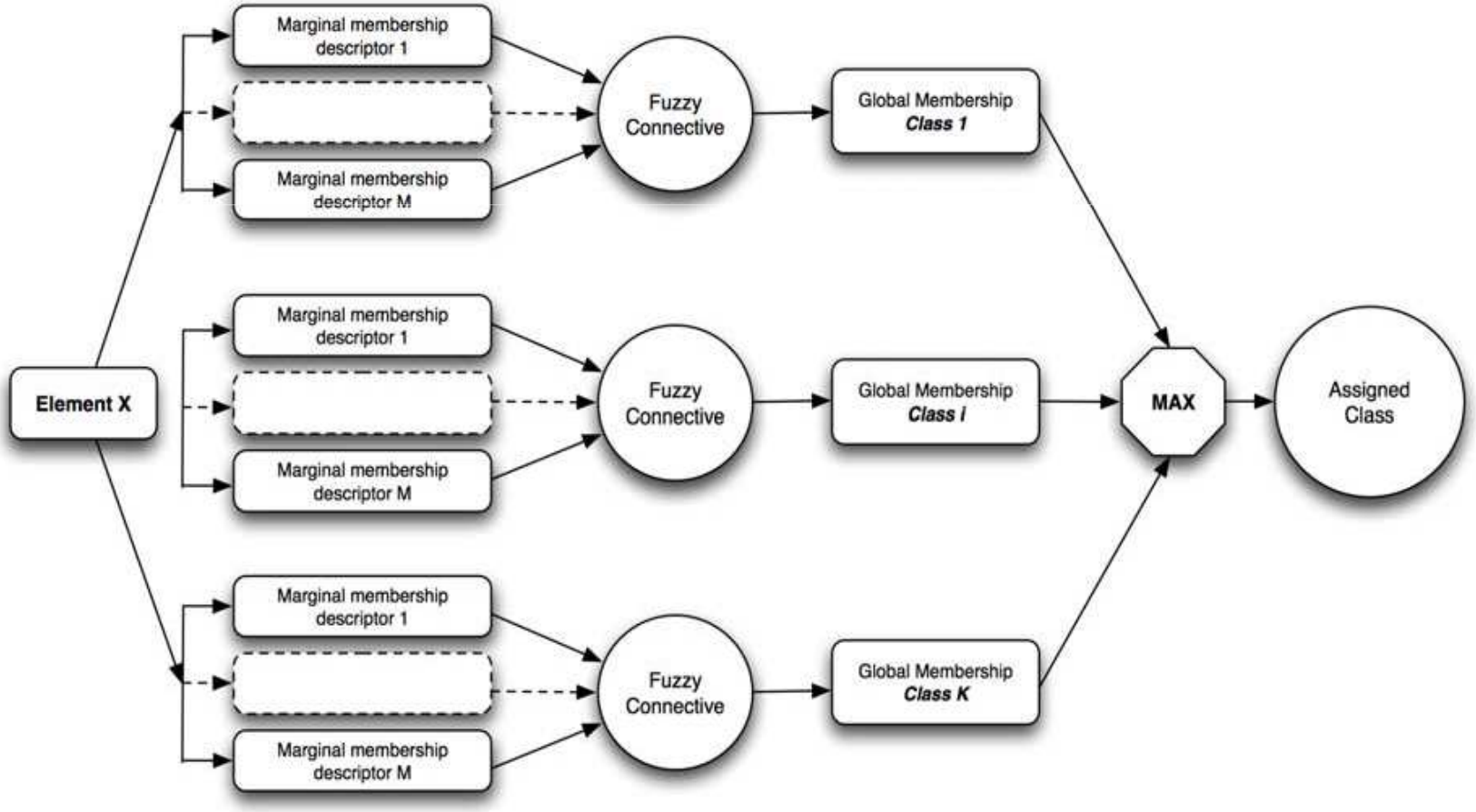
$$\mathrm{w}_{\mathrm{f}}^{*} = \frac{\mathrm{s}^{+}}{\| \mathrm{s}^{+} \|}$$

With

$$\mathrm{s} = \sum_{n=1}^{N} \min_{\{\tilde{c} \in C, \tilde{c} \neq C(\mathrm{x}_n)\}} \left\{ \mathrm{U}_{nc} - \mathrm{U}_{n\tilde{c}} \right\}$$

III Fuzzy learning algorithm

- Schematic illustration of LAMDA [J. Aguilar et al.1982]



IV. Marker selection for cancer prognosis

- **An accurate cancer prognosis is needed to help physicians select optimal treatment and reduce its related expensive medical costs.**
- **Usually, either clinical or genes markers are used to perform the prognosis.**
- **Integration of both information may improve the prognosis**
- **Two challenges are faced: High-dimension and heterogeneous data**

  - The first due to the presence of a large amount of irrelevant genes in microarray data

  - The second is related to the presence of mixed-type data (quantitative, qualitative and interval) in the clinical data

Experiments and Results

- **Pronostic du Métastase distant** : Netherlands Cancer Institute
  - 295 breast cancer patients
  - 29 patients with missing gene expression excluded from the study.
  - 2 classes according to the appearance of distant metastases: 88 patients with and 207 patients without.
  - Training D.(132) : 92 without, 40 with ; Test D. (134) : 93 without, 41 with.
  - Microarray dataset: 24188 gene expression
  - Clinical dataset: 10 variables – 1 of quantitative type; 1 intervallaire et 8 qualitatives
    - Age (quantitative)
    - Tumour grade (interval :[3,5] ; [6,7] ; [8,9])
    - Tumour size = T (qualitative: ≤2cm ; >2cm)
    - Nodal status = N (qualitative : pN0 ; '1-3' ; ≥4)
    - Mastectomy (qualitative : Yes, No)
    - Estrogen Receptor ER expression (qualitative : Yes, No)
    - Chemiotherapy (qualitative: Yes, No)
    - Hormonotherapy (qualitative: Yes, No)
    - St. Gallen - European criteria (qualitative: Chemio , No Chemio)
    - NIH –US criteria (qualitative: Chemio , No Chemio
    - Risk NIH (qualitative: low , intermediate , high)

Experiments and Results

- **Derived hybrid signature:** MEMBAS selects only 15 hybrid markers

  - Three are mixed-type clinical markers (Number of positive lymph nodes "qualitative" , ER "qualitative" and Grade "interval"), added to them 12 genes. (optimal Classif. Performance).

- **Comparatives results between hybrid, clinical, genetic signatures and classical clinical indices:**

	TP	FP	FN	TN	Sens.	Specif.	Accuracy
Hybrid	13	12	28	81	0.32	0.87	**94/134 (70.15%)**
70-genes	25	29	16	64	0.61	0.69	89/134 (66.42%)
Clinical	23	37	18	56	0.56	0.60	79/134 (58.96%)
NIH	41	91	0	2	1	0.02	41/134 (32.09%)
St Gallen	38	85	3	8	0.93	0.09	46/134 (34.33%)

- St. Gallen - Chimio recommandée quand un critère est vrai : ER negatif; ganglions positif; T>2cm ; Grade III ou II ; Age <35 ans.
- NIH: Chimio quand ganglions positifs ou Taille > 1cm

TP: True Positive ; FP: False Positive ; FN: False Negative ; TN: True Negative; Sens.: Sensitivity; Specif.:Specificity.

V Sensor selection methodology

- **The success of fault detection and diagnosis of complex process depends strongly on the selection of measurements that characterize accurately the process behavior .**

- **Large number of sensor increases the induced instrumentation cost and may degrades the diagnosis efficiency.**

- **Efficient sensor selection methodologies are required that:**

→ Monitor accurately and robustly fault detection in complex processes

→ While, assure a reduction in the instrumentation costs and improve the process safety and quality

V Sensor selection methodology

1) Fault identification using the fuzzy classification technique LAMDA (self-learning) (*)

2) Sensor selection based on MEMBAS method (*) (**)

3) Generation of behavioral pattern of the process based only on the selected set of sensors.

4) Online recognition and validation on unseen data.

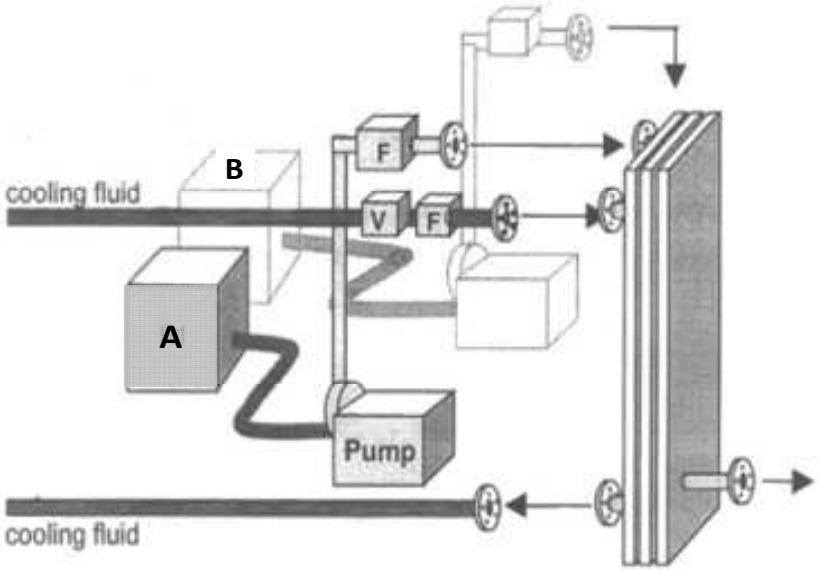
(*) Implemented on SALSA software tool [T. Kempowsky et al., 2003].

(**) Validated in an extensive experimental study on a large number of high dimensional and heterogeneous datasets [Hedjazi et al., 2010].

Application on chemical process

- **Pharmaceutical synthesis in a new intensified heat exchanger reactor equipped of 15 sensors: 12 internal temperatures (interval) , Utility outlet temperature (interval) , Reacts. Pressure (quantitative) .**

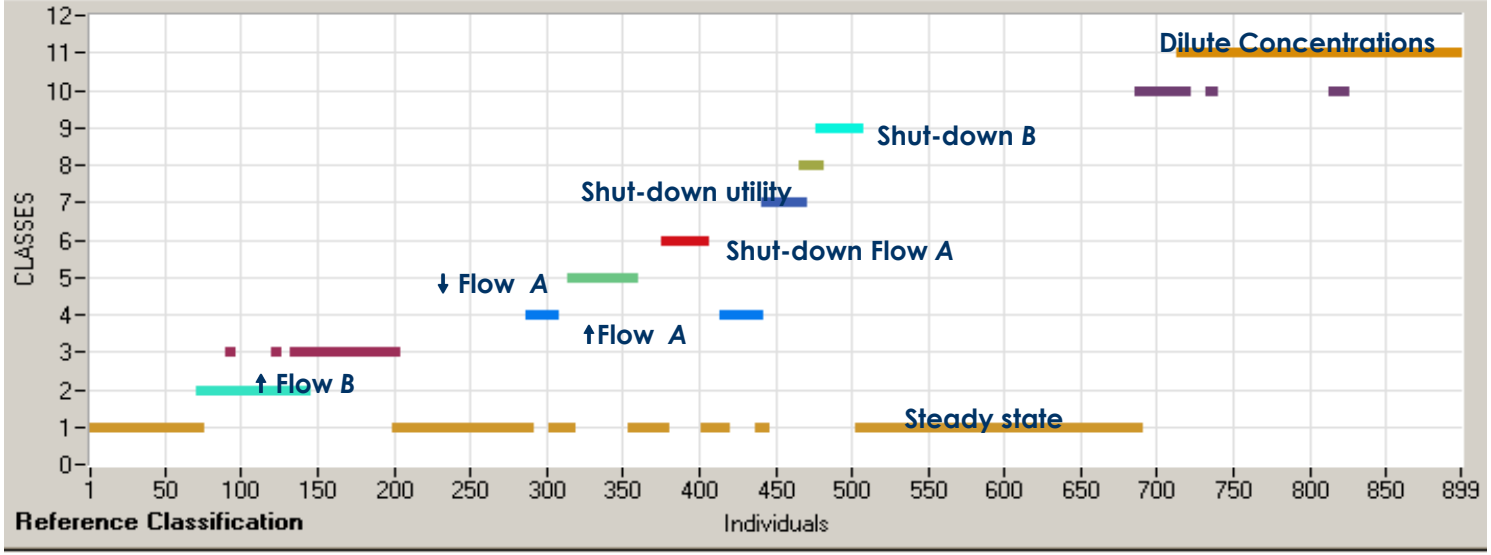
$$A_{(aq)} + B_{(l)} \rightarrow C_{(l)} + D_{(l)}$$
$$B_{(l)} + E_{(l)} \rightarrow 2\,D_{(l)}$$



Application on chemical process

1) Fault identification using the fuzzy clustering technique LAMDA.

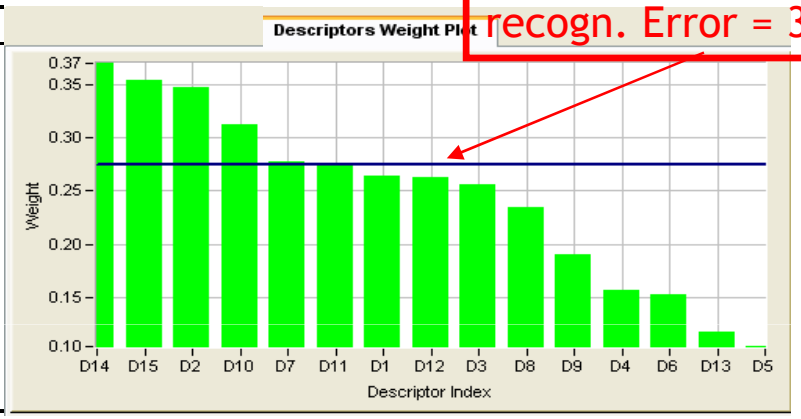
Class Number	Class Description
1	Steady state
2	Increased (⬆)flow reactant B
3	Critical Increased (⬆) flow reactant B
4	Increased (⬆) flow reactant A
5	Decreased (⬇) flow reactant A
6	Shut-down flow reactant A
7	Shut-down utility flow
8	Critical shut-down utility flow
9	Shut-down flow reactant B
10	Dilute concentration reactant A
11	Critical dilute concentration reactant A



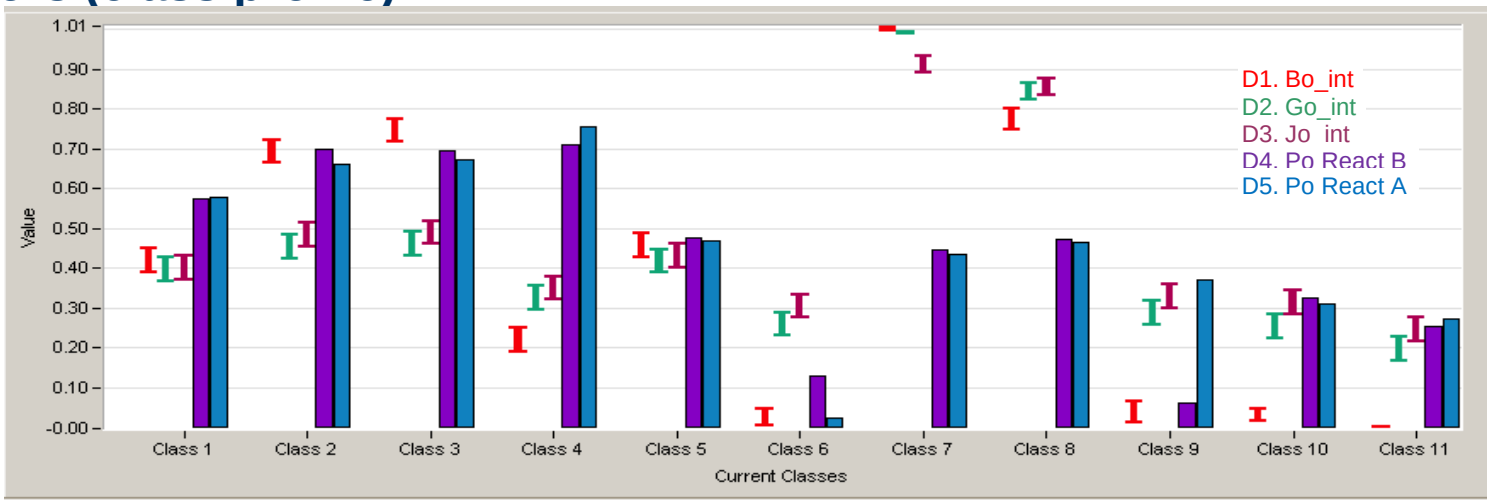
Application on chemical process

2) Sensor selection using MEMBAS method.

RANK	INDEX	WEIGHT	NAME
1	14	0.369726	P. Injection (B)
2	15	0.353165	P. Primary (A)
3	2	0.347069	Bo_int
4	10	0.312663	Jo_int
5	7	0.277623	Go_int
6	11	0.274547	Ko_int
7	1	0.264255	Ao_int
8	12	0.262668	Lo_int
9	3	0.25645	Co_int
10	8	0.234234	Ho_int



3) Generation of behavioral pattern of the process based on the selected sensors (class profile).

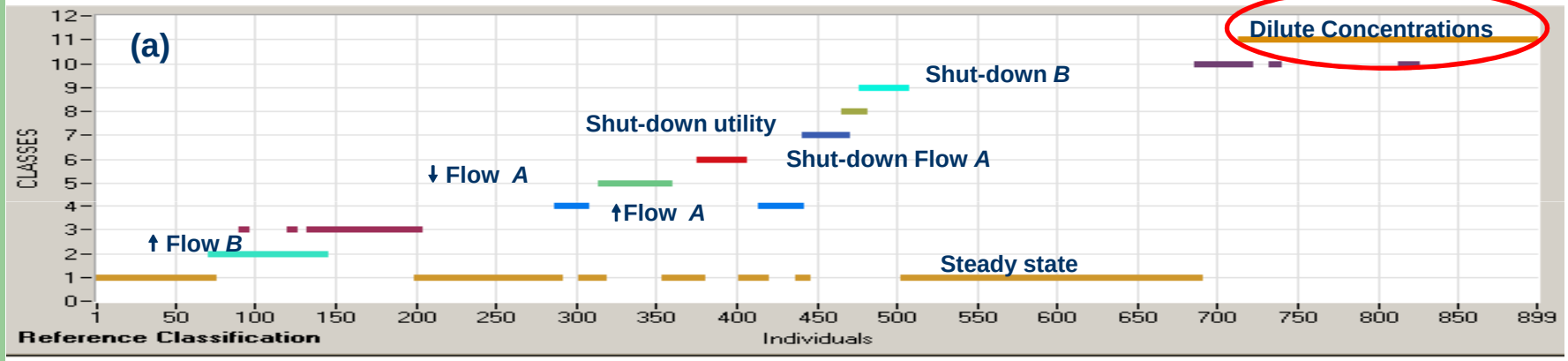


Application on chemical process

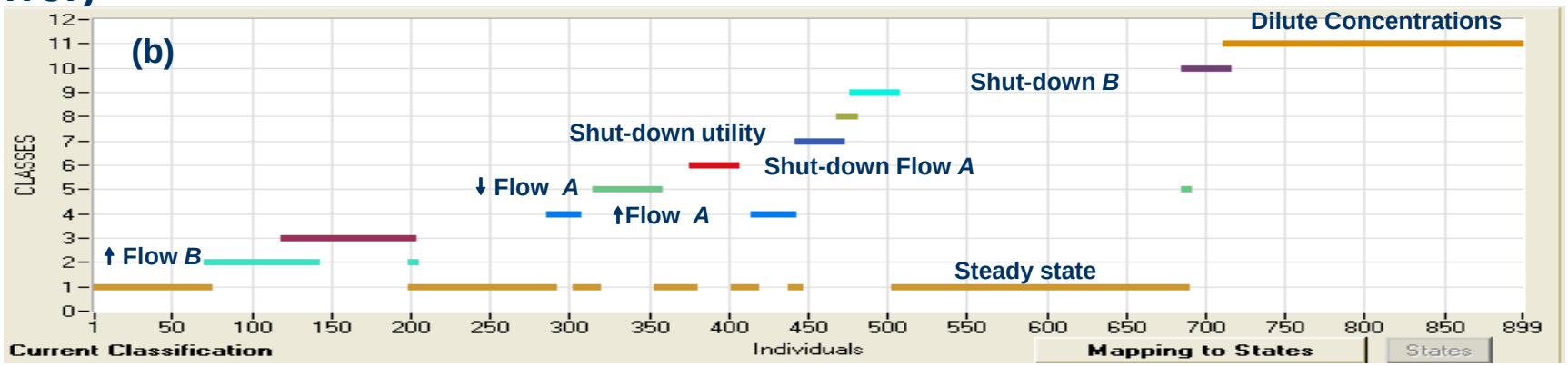
4) Online recognition and validation on unseen data

Fault on concentration detected only with temperature measurements

a) Faults identified using 15 original sensors



b) Recognition using 5 selected sensors by Membas (3.66% recognition error)



Application on chemical process

c) Validation on unseen data

Conclusion

1. Medical diagnosis:

↗ Two challenges faced for the integration of clinical and microarray data to perform cancer prognosis/diagnosis : High-dimensional and heterogeneous data.

↗ This approach can outperform classical approaches and selects meaningful hybrid markers signature: Three well known clinical markers (i.e. included in clinical indices) and twelve genes.

↗ Reduces significantly the number of markers needed to perform a cancer prognosis task (15 hybrid markers vs. 70 Amsterdam genes).

2. Industrial process diagnosis:

↗ Proposed approach handles interval data which are of big interest in practical situations to take into account inherent uncertainty to sensors measurement and noisy data (avoid false alarms).

↗ The proposed methodology is Data-driven based, does not require a physical model and is appropriate for Highly nonlinear and dimensional problems.

↗ Application on chemical process: High fault detection accuracy and reduced number of sensors (avoid expensive on-line concentration mea- surement)

Conclusion

↗ **Despite their behavioral difference, both domains industrial process and medical diagnosis exhibit many common practices:**

→ Sensor selection for industrial process diagnosis

→ Marker selection for medical diagnosis

↗ **A novel methodology enables to handle simultaneously both problems regardless of their own characteristics:**

→ Copes with the problem of high dimensionality based on classical optimization methods.

→ Handles appropriately heterogeneous data (quantitative, qualitative, interval)

→ Handling interval data which are of big interest in practical situations to take into account inherent uncertainty to sensors measurement and noisy data (avoid false alarms).