S-FRULER: aprendizaje automático escalable de reglas de predicción en Big Data

Ismael Rodríguez-Fernández, Manuel Mucientes, Alberto Bugarín
manuel.mucientes@usc.es

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Santiago de Compostela, SPAIN

citius.usc.es

Acknowledgements

- This work was supported by the Spanish Ministry of Economy and Competitiveness under project TIN2014-56633-C3-1-R, the Galician Ministry of Education under the projects EM2014/012 and CN2012/151, and Mestrelab Research S.L. under project IN852A 2013/75

CiTiUS

Motivation

- Genetic Fuzzy Systems (GFSs): fuzzy rules + evolutionary algorithms
  - Fuzzy rules: knowledge representation
  - Evolutionary algorithms: optimization
  - Accuracy + readability/interpretability

- TSK-1 fuzzy rules

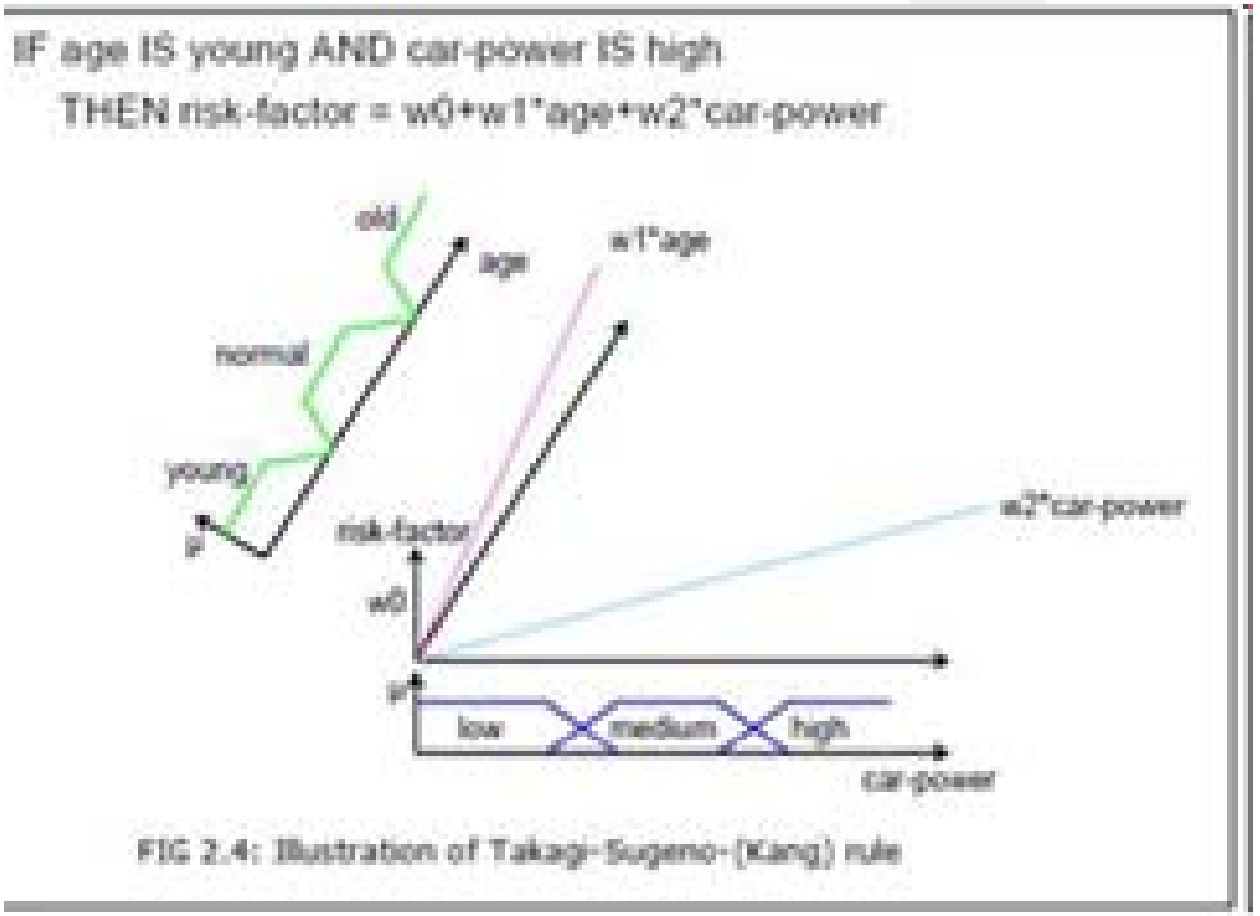


FIG 2.4: Illustration of Takagi-Sugeno-(Kang) rule

CITIUS

Motivation (ii)

- FRULER: accurate and simple TSK-1 fuzzy rule base models for regression

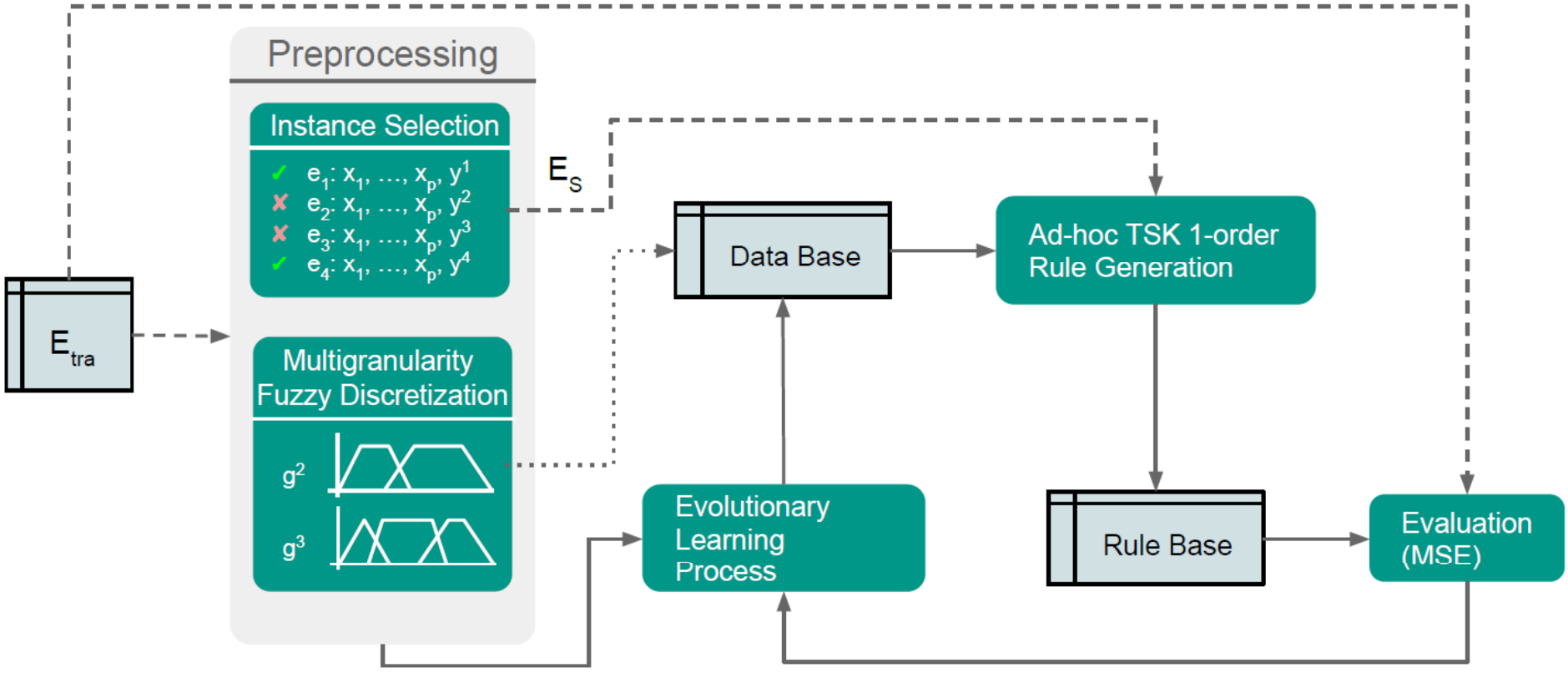
$$\text{If } X_1 \text{ is } A_1 \text{ and } X_2 \text{ is } A_2 \text{ and } \ldots \text{ and } X_p \text{ is } A_p \text{ then}$$
$$Y = \beta_0 + X_1 \cdot \beta_1 + X_2 \cdot \beta_2 + \cdots + X_p \cdot \beta_p$$

- Simplicity: improve readability/interpretability and generalization ability
  - Linguistic fuzzy partitions with a low number of labels
  - Low number of rules
  - Regularization of the consequents

- S-FRULER: scalable version of FRULER for Big Data

CiTIUS

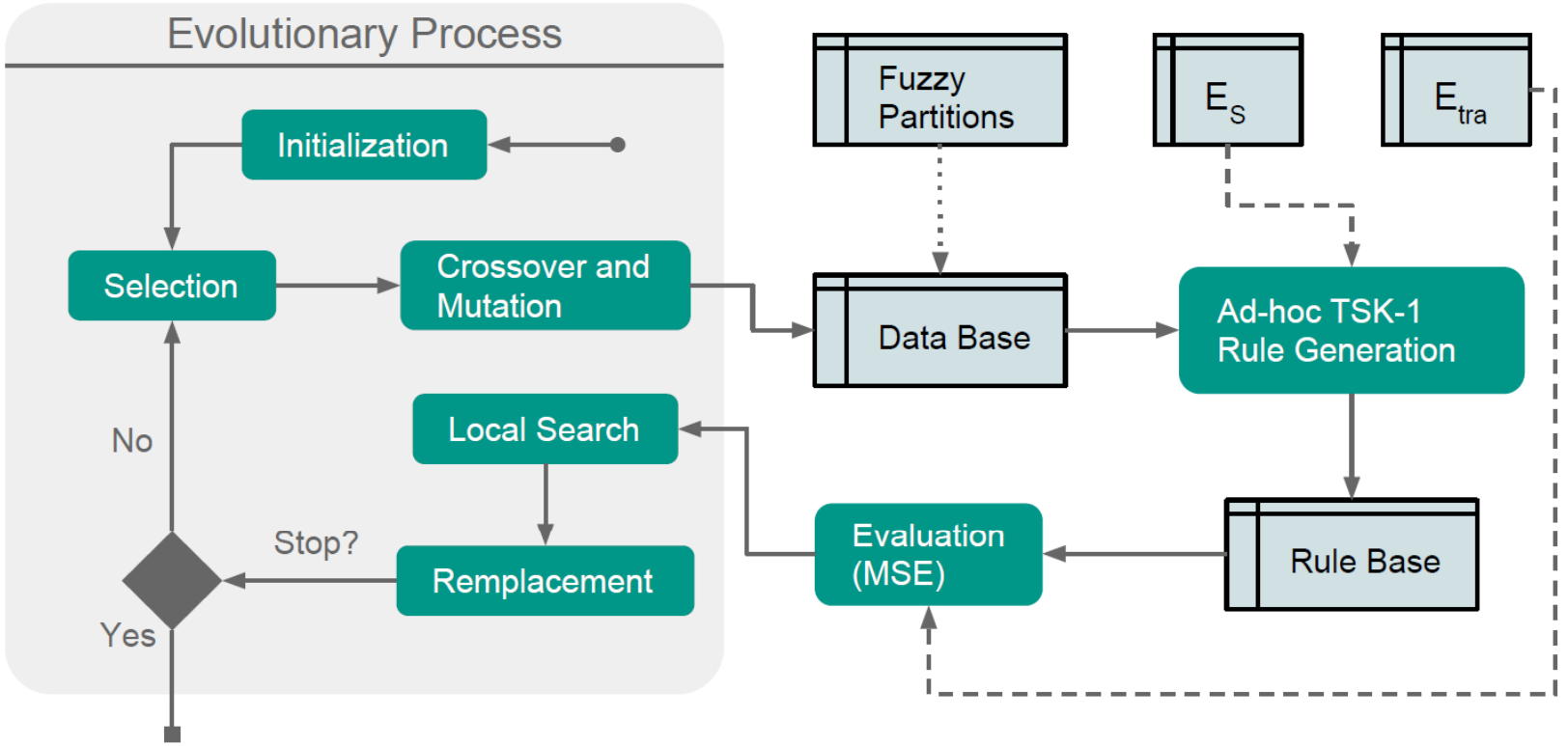
FRULER
Fuzzy Rule Learning through Evolution for Regression



CiTIUS

Genetic algorithm

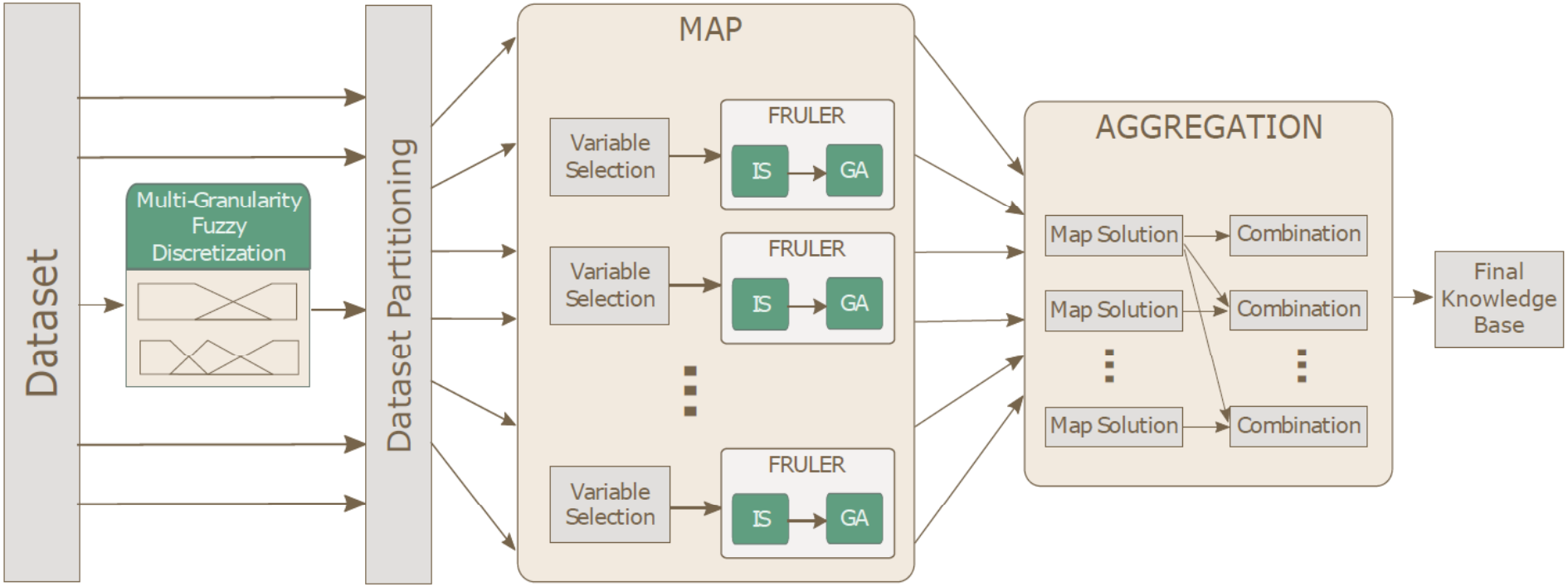
- Codification: $C_1 = (g_1, g_2, \ldots, g_{p_m})$ $C_2 = (\alpha_1^1, \ldots, \alpha_1^{g_1-1}, \ldots, \alpha_p^1, \ldots, \alpha_p^{g_p-1})$



- Generation of the KB: Wang & Mendel + Elastic Net (selected instances)
- Evaluation: training examples

S-FRULER

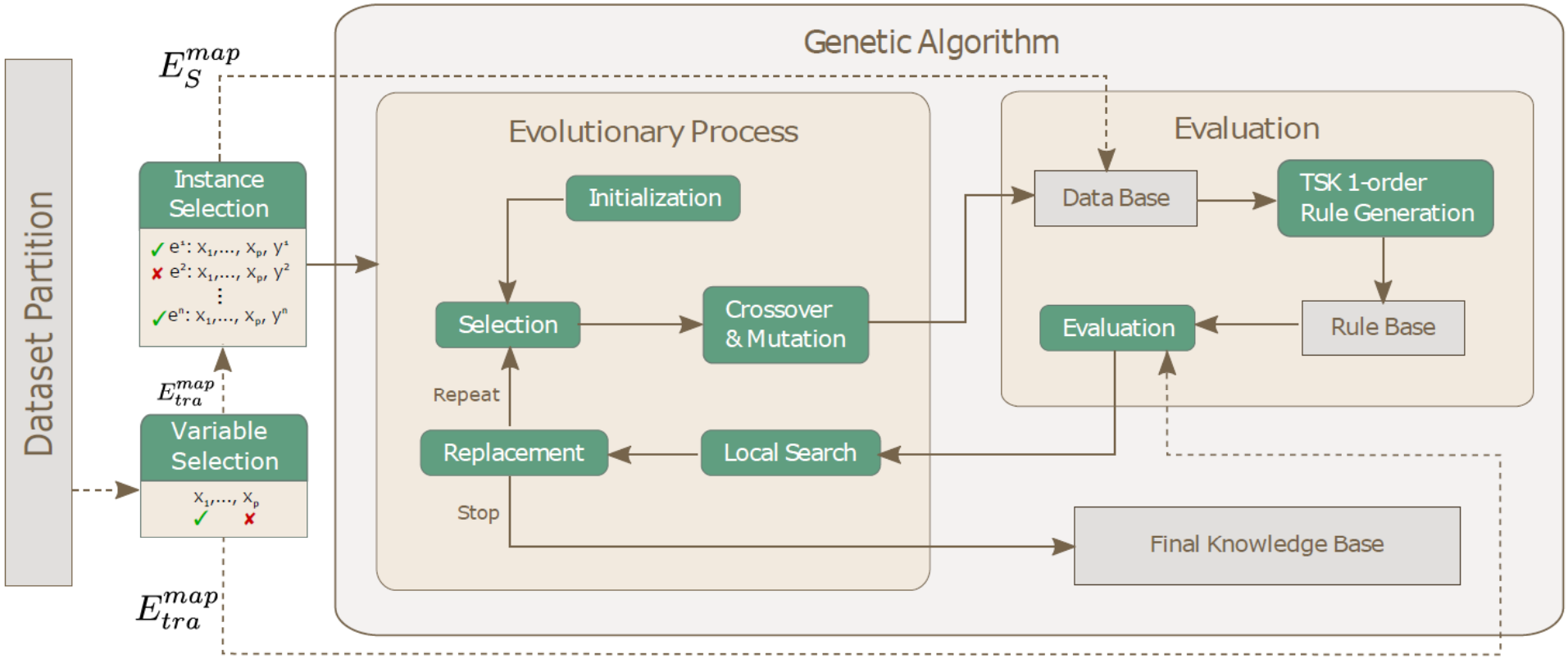
Scalable Fuzzy Rule Learning through Evolution for Regression



- Number of partitions: $n_{map} = log_2(p^2 * l * n)$
  - Depends on the number of input variables, the maximum granularity of the input variables, and the number of instances

CiTIUS

S-FRULER: Map function



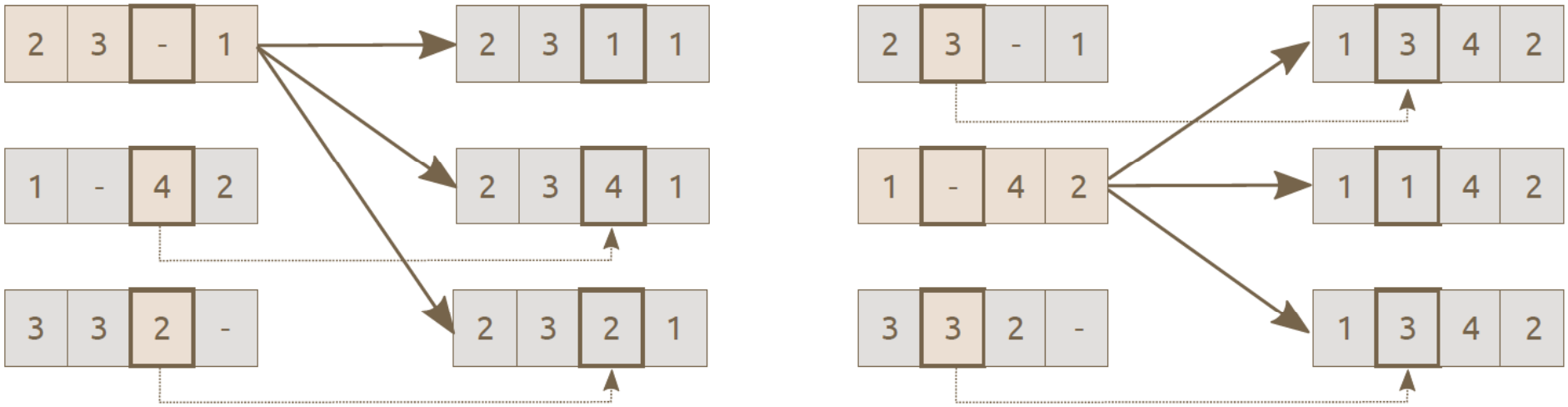
- For each node select randomly a subset of variables:
  - p_m: number of variables for each mapper

$$p_m \geq -p \cdot ((1 - \alpha_{p_m})^{1/n_{map}} - 1)$$

CiTIUS

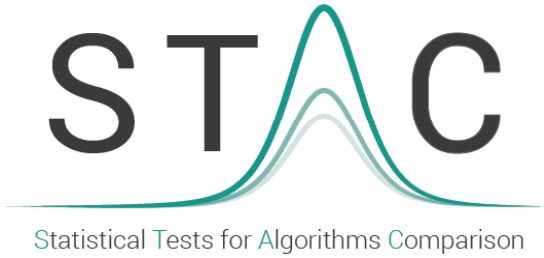
S-FRULER: aggregation function

- Keep the KB simple: combine the granularities of the different solutions
  - Maximum of n^2_{map} solutions
- Generate the TSK rule bases using the combination of selected instances



Results: FRULER

- 28 datasets
- **FRULER ranks first**



- http://tec.citius.usc.es/stac/

algorithms	FRULER		FS_{MOGFS}^{e}+TUN^{e}		L-METSK-HD^{e}		A-METSK-HD^{e}	
	#Rules	Test Error	#Rules	Test Error	#Rules	Test Error	#Rules	Test Error
ELE1	**4.1**	2.012	8.1	**1.954**	15	1.925	11.4	2.022
PLA	**1.4**	1.219	18.6	1.194	23	1.218	19.2	**1.136**
QUA	7.8	0.0181	**3.2**	**0.0178**	35.9	0.019	18.3	0.0181
ELE2	**4.3**	6,729	8	10,548	59	20,095	36.9	**3,192**
FRIE	**8.0**	**0.731**	22	3.138	95.1	3.084	66	1.888
MPG6	**13.7**	**3.727**	20	4.562	99.6	4.469	53.6	4.478
DELAIL	**2.5**	1.458	6.2	1.528	98.3	1.621	36.8	**1.402**
DEE	**7.9**	**0.080**	18.3	0.093	96.4	0.095	50.6	0.103
DELELV	**5.8**	1.045	7.9	1.086	91	1.119	39.1	1.031
ANA	**3.9**	0.008	10	**0.003**	48.9	0.006	33.3	0.004
MPG8	**12.7**	**4.084**	23	4.747	98.7	5.61	64.2	5.391
ABA	**4.5**	2.393	8	2.509	42.4	2.581	23.1	**2.392**
CON	**8.9**	**20.598**	15.4	32.977	96.5	38.394	53.7	23.885
STP	42.4	**0.353**	**23**	0.912	100	0.78	66.4	0.387
WAN	**5.6**	**0.888**	8	1.635	91.1	1.773	48	1.189
WIZ	**8.9**	**0.663**	10	1.011	55.4	1.296	29.1	0.944
FOR	**5.6**	**2,214**	10	2,628	93.7	4,633	40.6	5,587
MOR	7.9	**0.007**	7	0.019	40.9	0.028	27.2	0.013
TRE	**4.5**	**0.027**	9	0.044	42.8	0.052	28.1	0.038
BAS	**6.2**	305,777	17	**261,322**	95.7	320,133	59.8	368,820
CAL	15.4	2.110	**8.4**	2.95	99.8	2.638	55.8	**1.71**
MV	**6.0**	0.083	14	0.158	76.4	0.244	56.5	**0.061**
HOU	12.1	**8.005**	**11.7**	9.4	68.9	10.368	30.5	8.64
ELE	**5.4**	**2.934**	8	9	76.4	8.9	34.9	7.02
CA	**7.1**	**4.634**	14	5.216	71.3	5.88	32.9	4.949
POLE	40.8	110.898	**13.1**	102.816	100	150.673	46.3	**61.018**
PUM	**7.8**	0.367	17.6	0.292	87.5	0.594	63.3	**0.287**
AIL	**8.5**	**1.404**	15	2	99.1	1.822	48.4	1.51

CiTIUS

Results: FRULER for energy optimization

- Model the energy-building behavior and optimize energy efficiency
- OPERE(EU LIFE program)
- Monte da Condesa building:
  - 25,000 m^2
  - 2013 power consumption: 5,747 MWh
  - SCADA system with 469 variables
  - Dorm: 6 floors, > 400 students

	$\%R$		$\% < T_{setpoint}$		Temperature diff.
	oldSys	*newSys*	*oldSys*	*newSys*	Diff
07th to 10th of March	26.36%	16.67%	3.54%	3.33%	0.16 °C
05th to 08th of April	27.95%	10.42%	0.42%	0.21%	0.57 °C
04th to 07th of May	23.61%	0.00%	0.42%	0.00%	0.60 °C
02th to 05th of June	0.00%	1.04%	2.08%	0.00%	0.28 °C
Average	19.48%	7.03%	1.62%	0.89%	0.37 °C

Results: FRULER vs. S-FRULER

Dataset	FRULER	S-FRULER
DELAIL	1.46	**1.44**
DELELV	**1.04**	1.12
CAL	**2.11**	2.18
MV	0.08	**0.05**
HOU	**8.0**	8.2
ELV	**2.9**	3.2
CA	**4.6**	**4.6**
POLE	**111**	124
PUM	0.367	**0.349**
AIL	**1.4**	**1.4**

- "medium-sized" datasets: <40k examples, <40 variables
- FRULER best in 5, S-FRULER best in 3, equal in 2

CiTUS

Results: runtime of FRULER vs. S-FRULER

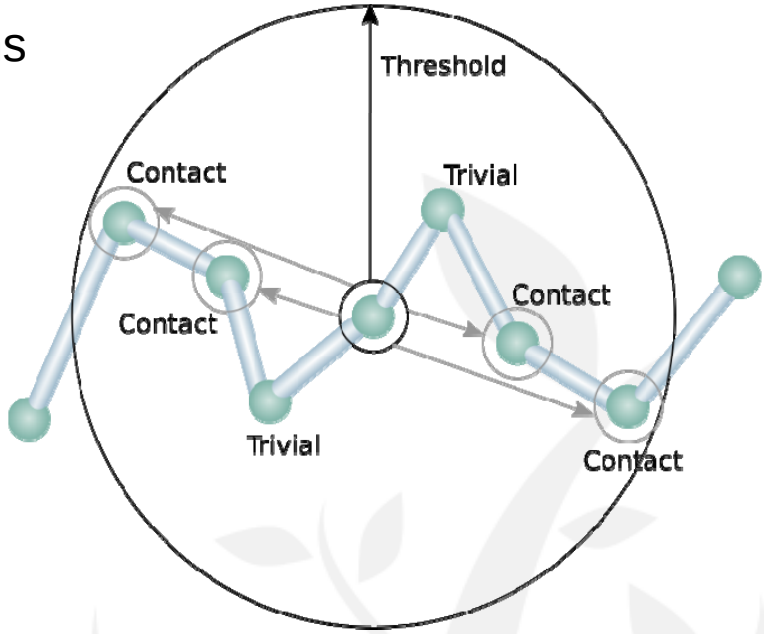
Algorithms	FRULER	S-FRULER	Standalone		Cluster	
	Time	n_{map}	Time	Speedup	Time	Speedup
DELAIL	0:09:58	21	0:01:18	8	0:00:48	12
DELELV	0:25:01	22	0:01:38	15	0:01:13	21
CAL	1:57:03	23	0:04:20	27	0:03:22	35
MV	1:17:02	23	0:09:27	8	0:05:49	13
HOU	4:15:17	25	0:04:06	62	0:03:09	81
ELV	3:01:30	26	0:03:10	57	0:03:14	56
CA	0:38:12	25	0:03:46	10	0:01:48	21
POLE	1:53:15	27	0:10:20	11	0:05:14	22
PUM	31:14:27	24	0:01:58	956	0:01:39	1,139
AIL	12:50:38	28	0:07:13	107	0:03:32	218

- Standalone: HP Proliant with 4 AMD Opteron 6262 HE (64 cores and 128 GB)
- Spark Cluster: Amazon Elastic MapReduce (EMR) 4.0.0 with m3.xlarge machines (Intel Xeon E5-2670 v2, 4 cores, 15 GB)

CiTIUS

Results: large dataset

- Prediction of the 3D structure of protein chains
  - Estimate the coordination number of a protein



- Regression: comparison with Mllib
  - Runtimes for standalone mode

			S-FRULER		Ridge SGD		Lasso SGD	
Datasets	# Cases	# Variables	Test Error	Time	Test Error	Time	Test Error	Time
w1	257,560	60	12.45	04:42:08	15.09	00:42:56	19.01	00:45:05
w2	257,560	100	12.15	05:32:33	14.20	01:00:44	18.91	01:01:25
w3	257,560	140	12.23	05:48:42	14.18	01:04:36	18.89	01:04:43
w4	257,560	180	12.30	12:17:42	13.62	01:06:42	18.93	01:05:46

CiTUS

Results: large dataset (ii)

- Classification:

Algorithm	S-FRULER	BioHEL	GAssist	PART	C4.5
w1	76.9	75.8	74.8	70.9	68.6
w2	77.3	76.0	74.7	70.9	68.6
w3	77.7	76.4	74.6	69.9	68.1
w4	77.4	76.5	74.8	76.0	68.2

CiTIUS

Chemical Shift Prediction

- Predict the chemical shift of the atoms in a molecule
- More than 54,000 examples, 985 input variables
- RF for regression

Error	< 0.1	< 0.2	< 0.3	< 0.4	< 0.5	< 0.6	< 0.7	< 0.8	< 0.9	< 1	> 1	> 2
Porcentaje de instancias	72.14	84.63	90.41	93.63	95.81	96.98	97.90	98.61	98.98	99.37	0.63	0.02
Desviación	0.03	0.05	0.07	0.09	0.10	0.11	0.13	0.14	0.14	0.15	0.27	0.00
% de instancias en 4 desviaciones	99.77	99.67	99.67	99.64	99.62	99.61	99.58	99.55	99.55	99.53	94.12	100

CiTiUS

Conclusions

- FRULER obtains simple models with high precision
  - Small and medium-sized datasets
- S-FRULER is a distributed version of FRULER implemented with Apache Spark
- S-FRULER has speedups usually larger than the number of dataset partitions used, showing an scalability higher than linear in both standalone and cluster modes
- Results demonstrate the capability of S-FRULER to obtain precise and simple models in large scale problems.

CiTIUS

I. Rodríguez-Fdez, M. Mucientes, and A. Bugarín. Fuzzy Rule Learning through Evolution for Regression. Information Sciences, 354:1-18, 2016.

I. Rodríguez-Fdez, M. Mucientes, and A. Bugarín. S-FRULER: Scalable fuzzy rule learning through evolution for regression. Knowledge-Based Systems, 110:255-266, 2016.

Software disponible en: http://tec.citius.usc.es/fruler/

Ismael Rodríguez-Fernández, Manuel Mucientes, Alberto Bugarín

manuel.mucientes@usc.es

Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Santiago de Compostela, SPAIN

citius.usc.es

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA CAMPUS VIDA

CiTIUS Centro Singular de Investigación en Tecnoloxías da Información