Clustering, prediction and ordinal classification of time series using machine learning techniques: applications

International PhD

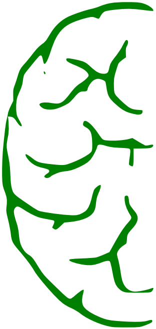
David Guijo Rubio
dguijo@uco.es
Supervisors:
César Hervás Martínez
Pedro Antonio Gutiérrez Peña

Dept. of Computer Science and Numerical Analysis, University of Córdoba.
Learning and Artificial Neural Networks (AYRNA) research group.

21st June 2021

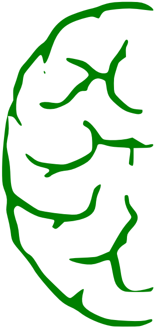
UNIVERSIDAD DE CÓRDOBA

Outline



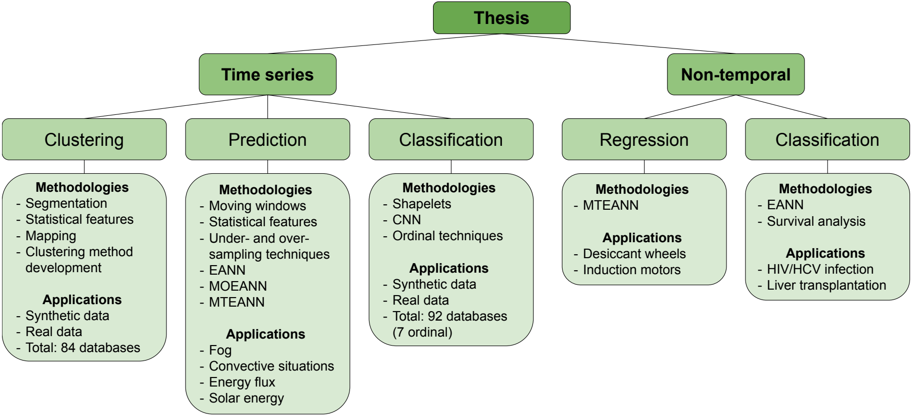
1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Outline



1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Clustering, prediction and ordinal classification of time series: applications

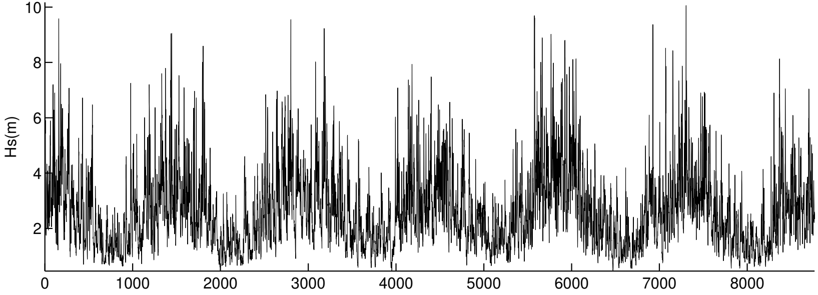


Outline

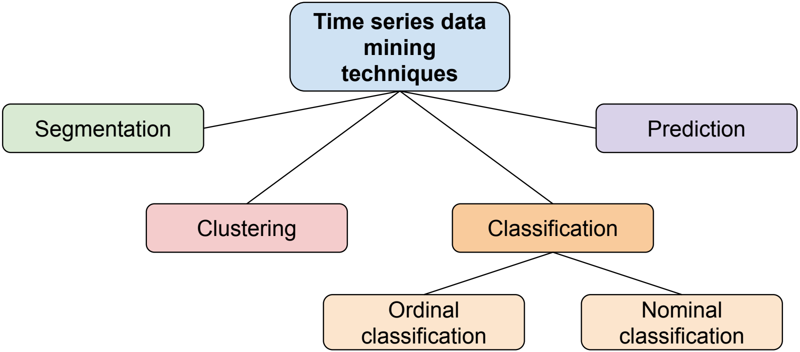
1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Time series data mining: **definition**

Time series are defined as temporal data collected chronologically or as a function varying across time.



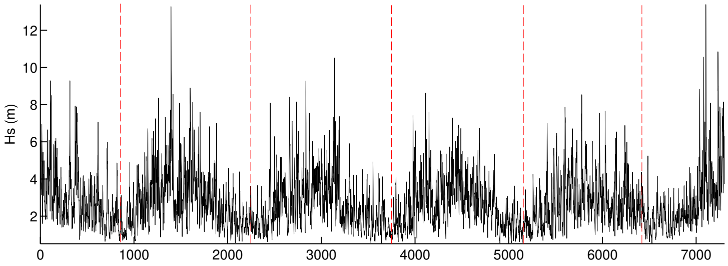
Time series data mining: **data mining techniques**



Others: regression, sequence discovery, ...

Time series data mining: **segmentation**

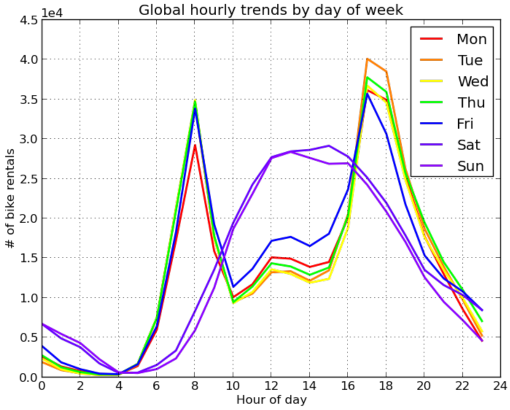
Given a time series $T = \{t_j\}_{j=1}^{N}$, the segmentation consists in finding m segments, defined by: $s_1 < s_2 < \ldots < s_{m-1}$.



According to the purpose of the analysis: grouping similar behaviours, obtaining an approximation (to simplify them), ...

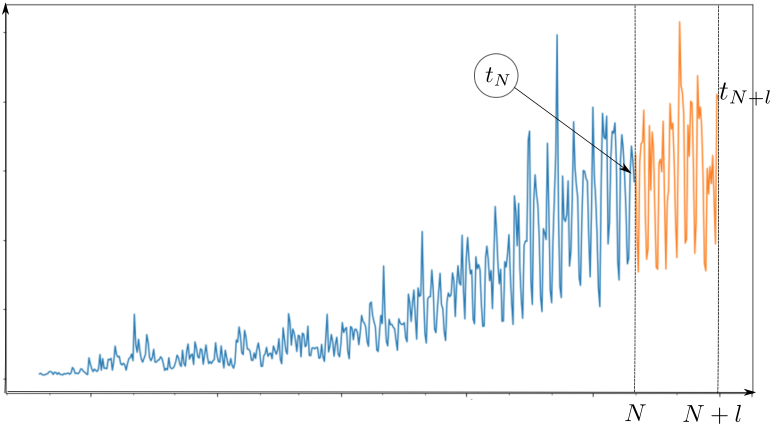
Time series data mining: **clustering**

Given a time series dataset $D = \{T_i\}_{i=1}^{M}$, the clustering consists in organising them into L groups, $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_L\}$.



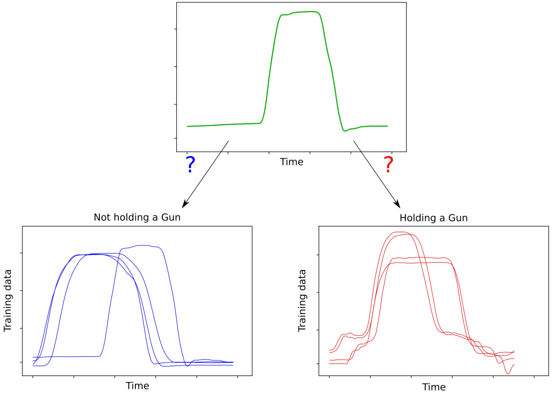
Time series data mining: **prediction**

Given a time series $T = \{t_j\}_{j=1}^{N}$, the prediction consists in the estimation of the value t_{N+l}.



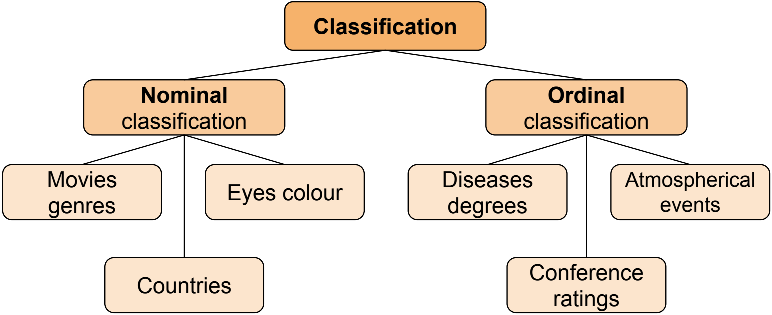
Time series data mining: **classification**

Given a time series $T = \{t_j\}_{j=1}^{N}$, the classification consists in identifying the class label y_j to which it belongs.



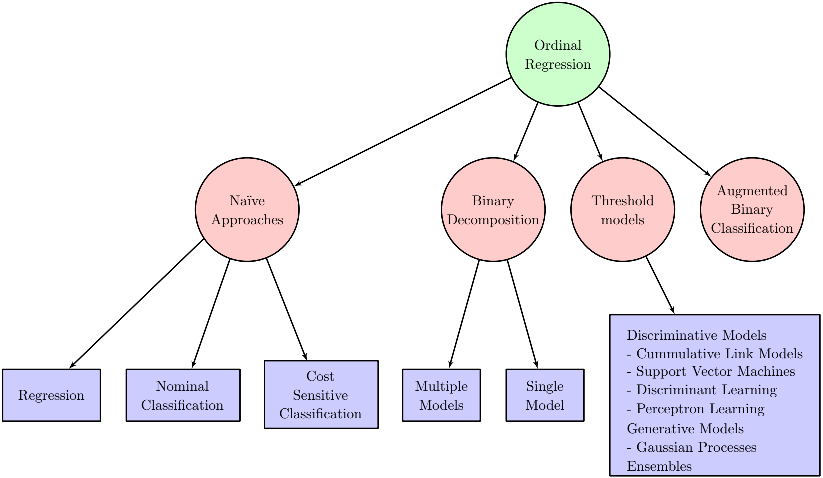
Time series data mining: **classification**

According to the nature of the labels assigned to the time series:

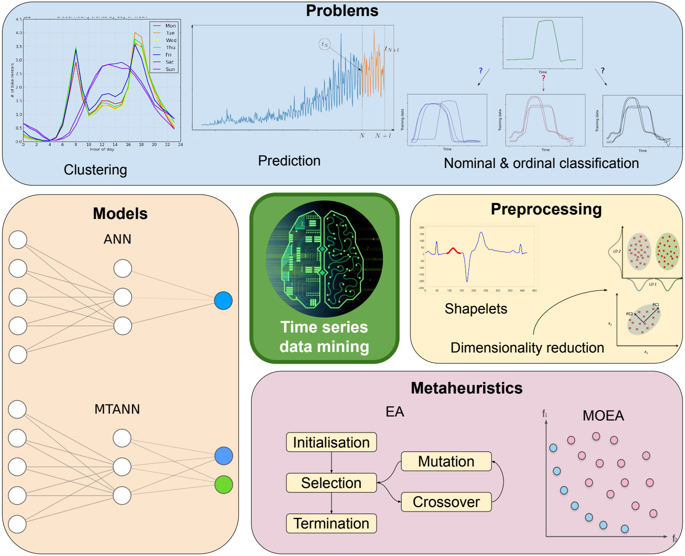


Time series data mining: **ordinal classification**

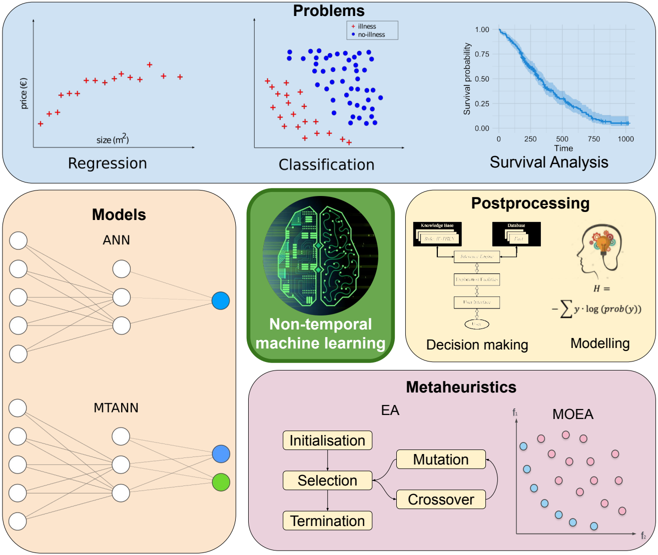
Taxonomy for ordinal classifiers:



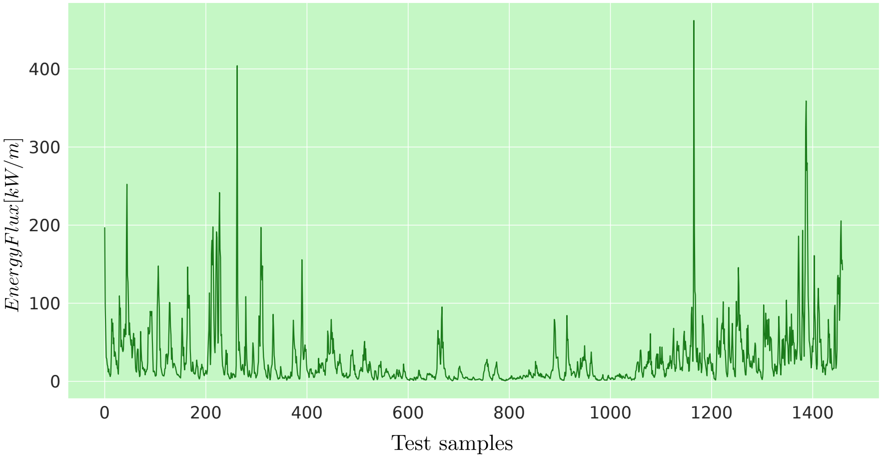
Methodologies for time series



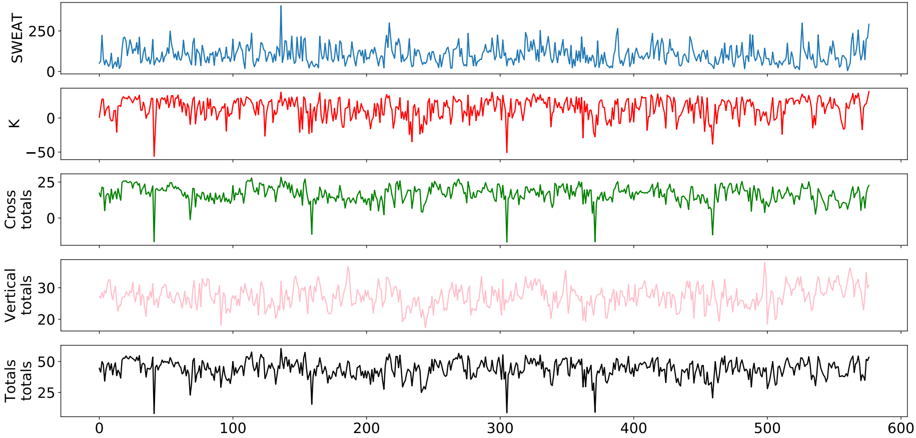
Methodologies for non-temporal data



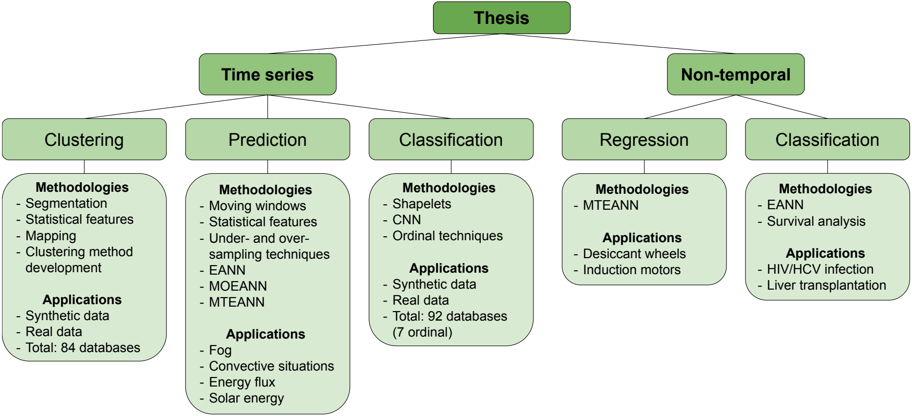
Real-world applications: **energy flux**



Real-world applications: **convective situations**



Clustering, prediction and ordinal classification of time series: applications



Outline

1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Objectives I

1. To propose different ANN architectures by hybridising activation functions or combining them in hidden and output layers.
2. To adapt EANNs for its application to two different sorts of problems: multi-objective and multi-task problems.
3. To review the state-of-the-art in preprocessing and analysis techniques for time series, with the aim of studying new representation forms alleviating the difficulty of subsequent tasks.
4. To study and develop a novel approach to time series clustering by preprocessing the time series with time series segmentation, reducing their dimensionality by carrying out a statistical feature extraction process.
5. To analyse and survey the ST methodology, in order to provide improvements to this methodology by developing a new proposal in the TSC field.

Objectives I

1. To propose different ANN architectures by hybridising activation functions or combining them in hidden and output layers.
2. To adapt EANNs for its application to two different sorts of problems: multi-objective and multi-task problems.
3. To review the state-of-the-art in preprocessing and analysis techniques for time series, with the aim of studying new representation forms alleviating the difficulty of subsequent tasks.
4. To study and develop a novel approach to time series clustering by preprocessing the time series with time series segmentation, reducing their dimensionality by carrying out a statistical feature extraction process.
5. To analyse and survey the ST methodology, in order to provide improvements to this methodology by developing a new proposal in the TSC field.

Objectives I

1. To propose different ANN architectures by hybridising activation functions or combining them in hidden and output layers.
2. To adapt EANNs for its application to two different sorts of problems: multi-objective and multi-task problems.
3. To review the state-of-the-art in preprocessing and analysis techniques for time series, with the aim of studying new representation forms alleviating the difficulty of subsequent tasks.
4. To study and develop a novel approach to time series clustering by preprocessing the time series with time series segmentation, reducing their dimensionality by carrying out a statistical feature extraction process.
5. To analyse and survey the ST methodology, in order to provide improvements to this methodology by developing a new proposal in the TSC field.

Objectives I

1. To propose different ANN architectures by hybridising activation functions or combining them in hidden and output layers.
2. To adapt EANNs for its application to two different sorts of problems: multi-objective and multi-task problems.
3. To review the state-of-the-art in preprocessing and analysis techniques for time series, with the aim of studying new representation forms alleviating the difficulty of subsequent tasks.
4. To study and develop a novel approach to time series clustering by preprocessing the time series with time series segmentation, reducing their dimensionality by carrying out a statistical feature extraction process.
5. To analyse and survey the ST methodology, in order to provide improvements to this methodology by developing a new proposal in the TSC field.

Objectives I

1. To propose different ANN architectures by hybridising activation functions or combining them in hidden and output layers.
2. To adapt EANNs for its application to two different sorts of problems: multi-objective and multi-task problems.
3. To review the state-of-the-art in preprocessing and analysis techniques for time series, with the aim of studying new representation forms alleviating the difficulty of subsequent tasks.
4. To study and develop a novel approach to time series clustering by preprocessing the time series with time series segmentation, reducing their dimensionality by carrying out a statistical feature extraction process.
5. To analyse and survey the ST methodology, in order to provide improvements to this methodology by developing a new proposal in the TSC field.

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
   1. Prediction of fog formation in airports.
   2. Prediction of convective situations formation in airports.
   3. Prediction of solar radiation.
   4. Prediction of energy flux from ocean waves.
   5. Modelling of Desiccant Wheels (DW).
   6. Modelling of the acoustic behaviour of induction motors.
   7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
   8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
    1. Prediction of fog formation in airports.
    2. Prediction of convective situations formation in airports.
    3. Prediction of solar radiation.
    4. Prediction of energy flux from ocean waves.
    5. Modelling of Desiccant Wheels (DW).
    6. Modelling of the acoustic behaviour of induction motors.
    7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
    8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
    1. Prediction of fog formation in airports.
    2. Prediction of convective situations formation in airports.
    3. Prediction of solar radiation.
    4. Prediction of energy flux from ocean waves.
    5. Modelling of Desiccant Wheels (DW).
    6. Modelling of the acoustic behaviour of induction motors.
    7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
    8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
   1. Prediction of fog formation in airports.
   2. Prediction of convective situations formation in airports.
   3. Prediction of solar radiation.
   4. Prediction of energy flux from ocean waves.
   5. Modelling of Desiccant Wheels (DW).
   6. Modelling of the acoustic behaviour of induction motors.
   7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
   8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
    1. Prediction of fog formation in airports.
    2. Prediction of convective situations formation in airports.
    3. Prediction of solar radiation.
    4. Prediction of energy flux from ocean waves.
    5. Modelling of Desiccant Wheels (DW).
    6. Modelling of the acoustic behaviour of induction motors.
    7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
    8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
    1. Prediction of fog formation in airports.
    2. Prediction of convective situations formation in airports.
    3. Prediction of solar radiation.
    4. Prediction of energy flux from ocean waves.
    5. Modelling of Desiccant Wheels (DW).
    6. Modelling of the acoustic behaviour of induction motors.
    7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
    8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
    1. Prediction of fog formation in airports.
    2. Prediction of convective situations formation in airports.
    3. Prediction of solar radiation.
    4. Prediction of energy flux from ocean waves.
    5. Modelling of Desiccant Wheels (DW).
    6. Modelling of the acoustic behaviour of induction motors.
    7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
    8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
    1. Prediction of fog formation in airports.
    2. Prediction of convective situations formation in airports.
    3. Prediction of solar radiation.
    4. Prediction of energy flux from ocean waves.
    5. Modelling of Desiccant Wheels (DW).
    6. Modelling of the acoustic behaviour of induction motors.
    7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
    8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
    1. Prediction of fog formation in airports.
    2. Prediction of convective situations formation in airports.
    3. Prediction of solar radiation.
    4. Prediction of energy flux from ocean waves.
    5. Modelling of Desiccant Wheels (DW).
    6. Modelling of the acoustic behaviour of induction motors.
    7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
    8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Objectives II

6. To adapt and develop a novel approach based on the ST technique for its application to ordinal data, opening a new branch in TSC known as TSOC.
7. To apply the methods described above to the following real-world problems belonging to national research projects[1]:
   1. Prediction of fog formation in airports.
   2. Prediction of convective situations formation in airports.
   3. Prediction of solar radiation.
   4. Prediction of energy flux from ocean waves.
   5. Modelling of Desiccant Wheels (DW).
   6. Modelling of the acoustic behaviour of induction motors.
   7. Identification of Human Immunodeficiency Virus/Hepatitis C Virus (HIV/HCV) co-infected patient typology.
   8. Donor-recipient matching in Liver Transplantation (LT).

[1] Collaboration with the University of Alcalá de Henares

Publications derived from the Thesis

11 papers in international journals

- 7 papers published in journals indexed in JCR (Q1).
- 3 papers published in journals indexed in JCR (Q2).
- 1 paper published in a journal indexed in JCR (Q3).

15 papers in conferences

- 11 papers published in international conferences.
- 4 papers published in national conferences.

Outline

1. Introduction
2. Background
3. Objectives
4. **Time series clustering**
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Time series clustering

D. Guijo-Rubio, A.M. Durán-Rosal, P.A. Gutiérrez, A. Troncoso and C. Hervás-Martínez. "Time series clustering based on the characterisation of segment typologies", IEEE Transactions on Cybernetics. 2020. JCR (2019): 11.079 Position: 5/136 (Q1).

Problem

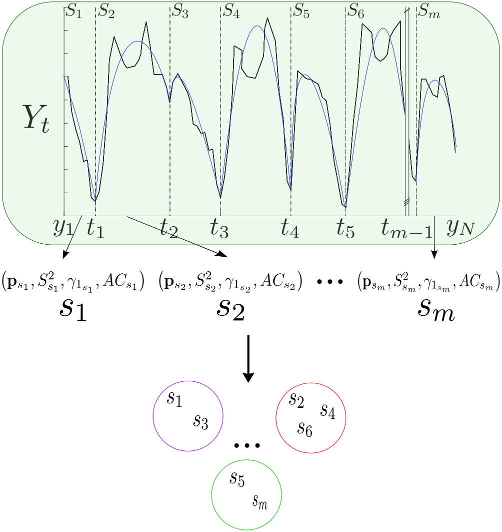
- We aim to group time series with respect to their similarity or characteristics.
- Clustering huge time series datasets with long time series are computationally intensive.
- Most of the clustering techniques use specific distance measures for time series.

Time series clustering

Methodology

- The methodology is composed of two clustering steps:
  1. Applied to each time series, and acting as dimensionality reduction:
     - Time series segmentation.
     - Segments mapping to equal length.
     - Segments clustering.
  2. Applied to the whole dataset to discover the groups of time series:
     - Time series mapping to common representation.
     - Time series clustering.
     - Clustering quality measurement.
- Designed for huge time series datasets with long time series.

Methodology I



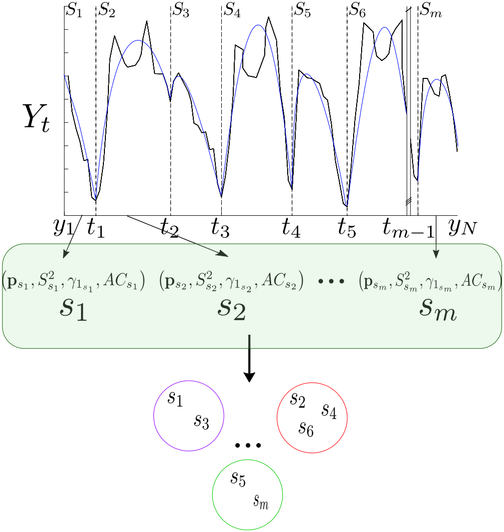
Time series segmentation: SwiftSeg has been applied. It introduces points into a growing window until an error threshold is exceeded:

$$SEP_s = \frac{\sqrt{SSE_s}}{|\overline{Y_s}|}, \quad (1)$$

$$SSE_s = \sum_{i=t_{s-1}}^{t_s} (\hat{y}_i - y_i)^2, \quad (2)$$

$$\overline{Y_s} = \frac{1}{t_s - t_{s-1} + 1} \sum_{i=t_{s-1}}^{t_s} y_i, \quad (3)$$

Methodology I



Segments mapping: each segment is projected into:

$$\mathbf{v}_s = (\mathbf{p}_s, S_s^2, \gamma_{1_s}, AC_s)$$

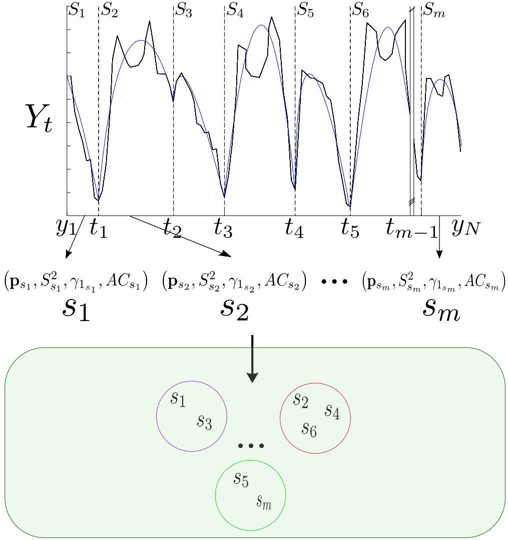
$\mathbf{p}_s$: parameters of the polynomial approximation

$$S_s^2 = \frac{1}{t_s - t_{s-1} + 1} \sum_{i=t_{s-1}}^{t_s} (y_i - \overline{y_s})^2 \tag{4}$$

$$\gamma_{1s} = \frac{\frac{1}{t_s - t_{s-1} + 1} \sum_{i=t_{s-1}}^{t_s} (y_i - \overline{y_s})^3}{\hat{\sigma}_s^3} \tag{5}$$

$$AC_s = \frac{\sum_{i=t_{s-1}}^{t_s} (y_i - \overline{y_s}) \cdot (y_{i+1} - \overline{y_s})}{S_s^2} \tag{6}$$

Methodology I



Segments clustering: a hierarchical clustering has been applied to the segments of the time series. Goals:

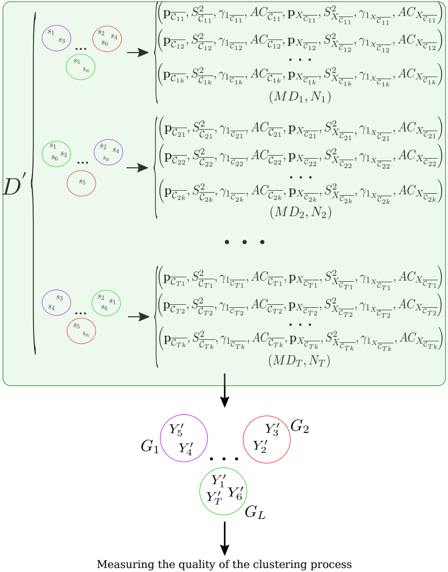
- Representing time series with the same length.
- Reducing the size of the time series significantly.

Clustering details: Agglomerative + Ward distance + $k = 2$.

Methodology II

Time series mapping: to represent all time series in the same dimensional space. For this:

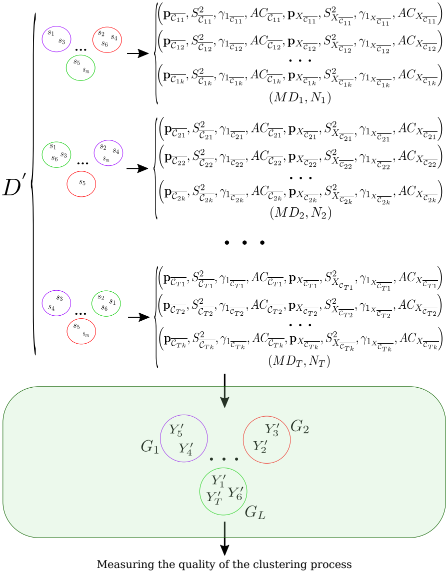
- $(\overline{\mathcal{C}_{ij}}, X_{\mathcal{C}_{ij}})\ \forall i \in \{1, \ldots, T\}, \forall j \in \{1, \ldots, k\}$
- The error difference $(MD_{\mathcal{C}_i})$ between the farthest segment and the closest segment.
- The number of segments of the time series, $N_{\mathcal{C}_i}$.



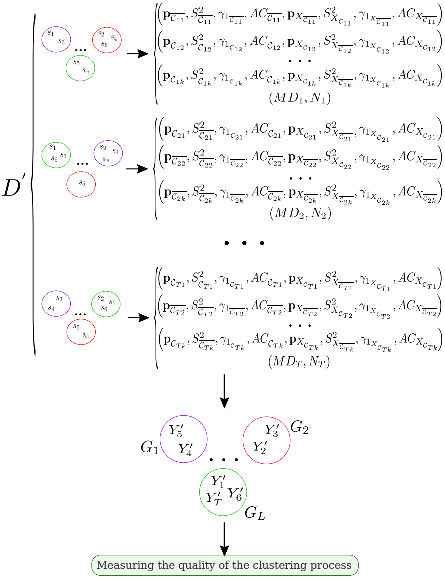
Methodology II

Time series clustering:

- Agglomerative hierarchical methodology.
- Ward distance.
- The number of clusters, L, is defined as the number of classes of the dataset.



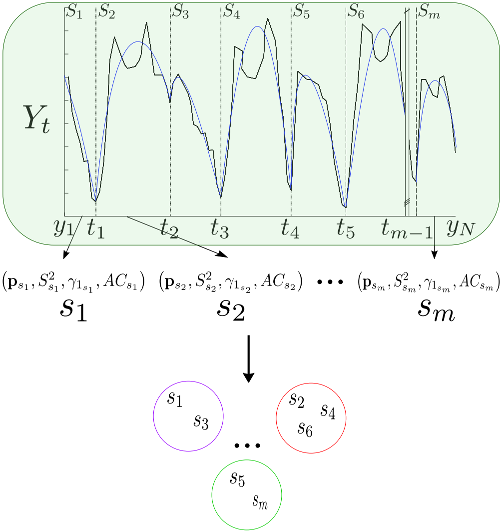
Methodology II



Clustering quality measurement:

- Internal measures: SSE, CH, SI, DB, DU, COP
- External measures: RI

Parameter adjustment



Parameter adjustment for the segmentation:

- TS3C$_{CH}$ $\rightarrow$ Selecting the $SEP_{\max}$ leading to the best Calinski and Harabasz index (CH).
- TS3C$_{MV}$ $\rightarrow$ Selecting the $SEP_{\max}$ which obtains the best value for the highest number of internal measures.

Results

Results

- 84 datasets from the UEA/UCR TSC.
- Comparison against 3 time series clustering techniques: DD_{DTW}-HC, KSC y WDTW.

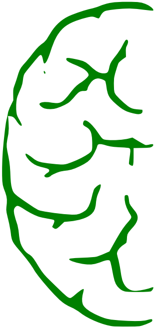
NumSeries	Length	Ranking RI		Ranking Time	
> 200(71)	> 300(34)	TS3C_{CH}	**2.529**	TS3C_{CH}	**1.824**
		TS3C_{MV}	2.779	TS3C_{MV}	2.824
		DD_{DTW}-HC	3.677	DD_{DTW}-HC	5.000
		KSC	3.427	KSC	2.706
		WDTW	*2.588*	WDTW	*2.647*
	< 300(37)	TS3C_{CH}	3.149	TS3C_{CH}	2.703
		TS3C_{MV}	*2.960*	TS3C_{MV}	3.703
		DD_{DTW}-HC	3.054	DD_{DTW}-HC	5.000
		KSC	3.122	KSC	*2.405*
		WDTW	**2.716**	WDTW	**1.189**
< 200(13)	All	TS3C_{CH}	3.423	TS3C_{CH}	3.000
		TS3C_{MV}	3.423	TS3C_{MV}	4.000
		DD_{DTW}-HC	3.308	DD_{DTW}-HC	5.000
		KSC	**2.385**	KSC	**1.385**
		WDTW	*2.462*	WDTW	*1.615*

Conclusions

Conclusions

- Significant differences for group 1, specially considering computational time.
- An important time series reduction performed.
- Lead to competitive clustering quality and computational time for large datasets.

Outline

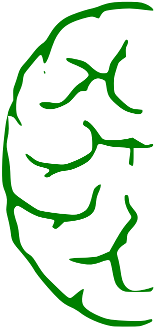


1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Introduction

- Time series prediction is usually accomplished by considering standard statistical procedures.
- We propose to use higher levels of representation of the time series.
- Then, we transform the prediction problems into ML tasks:
  - Time series ordinal prediction.
  - Time series forecasting.

Outline



1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Prediction of low-visibility events using ordinal classification

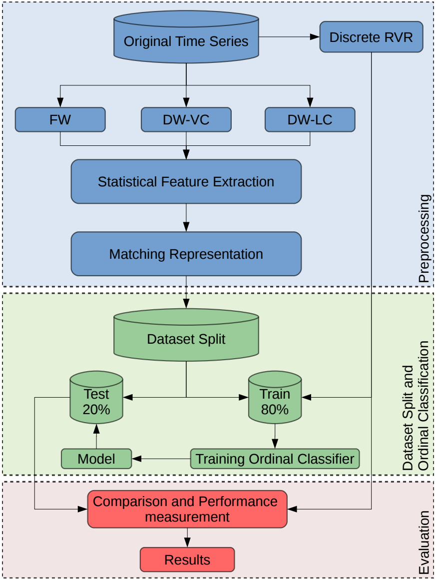
D. Guijo-Rubio, P.A. Gutiérrez, C. Casanova-Mateo, J. Sanz-Justo, S. Salcedo-Sanz and C. Hervás-Martínez. "Prediction of low-visibility events due to fog using ordinal classification", Atmospheric Research, Vol. 214, 2018, pp. 64 – 73. JCR (2018): 4.114. Position: 13/86 (Q1).

Problem

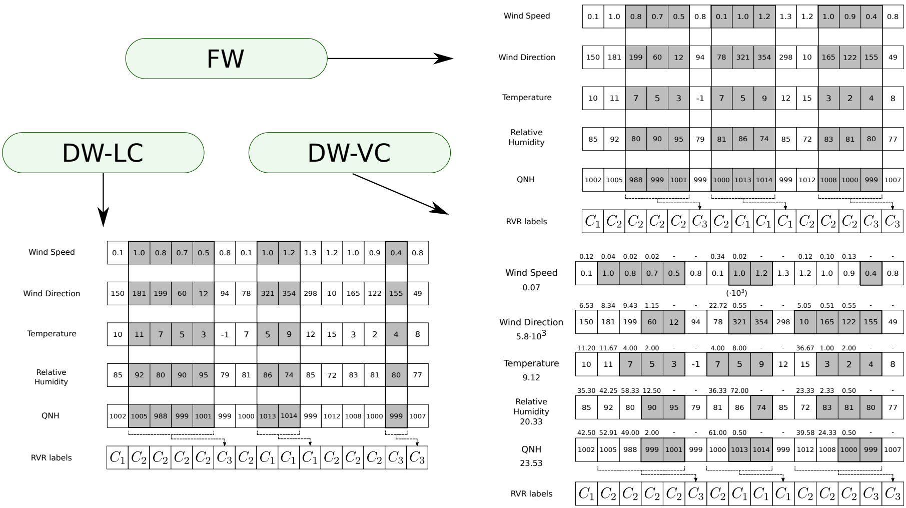
- The prediction of low-visibility events is crucial in transportation facilities such as airports, where they can cause severe impact in flight scheduling and safety.
- These events are characterised by the Runway Visual Range variable.

Methodology + Dataset

Class of day (C_q)	Training	Test
FOG	18 (3%)	15 (9%)
MIST	201 (30%)	59 (35%)
CLEAR	447 (67%)	94 (56%)
Total	666 (80%)	168 (20%)



Windows



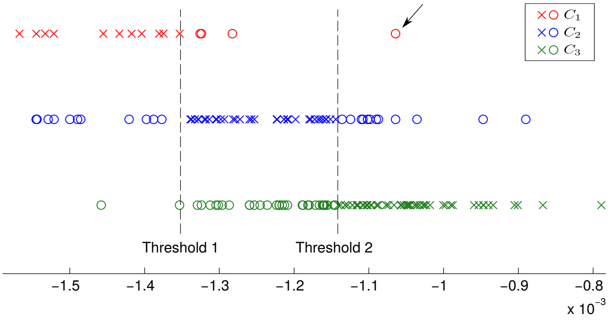
FW
DW-LC
DW-VC
Wind Speed
0.1 1.0 0.8 0.7 0.5 0.8 0.1 1.0 1.2 1.3 1.2 1.0 0.9 0.4 0.8
Wind Direction
150 181 199 60 12 94 78 321 354 298 10 165 122 155 49
Temperature
10 11 7 5 3 -1 7 5 9 12 15 3 2 4 8
Relative Humidity
85 92 80 90 95 79 81 86 74 85 72 83 81 80 77
QNH
1002 1005 988 999 1001 999 1000 1013 1014 999 1012 1008 1000 999 1007
RVR labels
C_1 C_2 C_2 C_2 C_2 C_3 C_2 C_1 C_1 C_1 C_2 C_2 C_2 C_3 C_3
Wind Speed 0.07
0.12 0.04 0.02 0.02 - - 0.34 0.02 - - 0.12 0.10 0.13 - -
($\cdot 10^3$)
Wind Direction $5.8 \cdot 10^3$
6.53 8.34 9.43 1.15 - - 22.72 0.55 - - 5.05 0.51 0.55 - -
Temperature 9.12
11.20 11.67 4.00 2.00 - - 4.00 8.00 - - 36.67 1.00 2.00 - -
Relative Humidity 20.33
35.30 42.25 58.33 12.50 - - 36.33 72.00 - - 23.33 2.33 0.50 - -
QNH 23.53
42.50 52.91 49.00 2.00 - - 61.00 0.50 - - 39.58 24.33 0.50 - -

Results

	AMAE (↓)						
	Type of window						
Predictor	FW	DWLC	DWVC	FW +DWLC	FW +DWVC	DWLC + DWVC	FW+DWLC +DWVC
SVR	0.585	0.558	0.549	0.525	0.590	*0.524*	**0.523**
POM	0.567	**0.521**	0.569	*0.524*	0.554	0.608	0.566
SVOREX	0.575	*0.541*	0.670	**0.526**	0.610	0.573	0.570
SVORIM	0.571	*0.542*	0.670	**0.528**	0.610	0.562	0.575
KDLOR	0.485	*0.382*	0.569	0.384	0.506	0.449	**0.369**
Persistence	0.434						

	MS (↑)						
	Type of window						
Predictor	FW	DWLC	DWVC	FW +DWLC	FW +DWVC	DWLC + DWVC	FW+DWLC +DWVC
SVR	0.000	0.000	0.000	0.000	0.000	0.000	0.000
POM	0.000	0.000	0.000	0.000	**6.666**	0.000	0.000
SVOREX	0.000	0.000	**13.333**	0.000	**13.333**	0.000	0.000
SVORIM	0.000	0.000	**13.333**	0.000	**13.333**	0.000	0.000
KDLOR	44.068	38.983	37.288	*52.542*	37.288	44.068	**61.017**
Persistence	42.857						

Conclusions



KDLOR

- The combined use of nonlinear classifiers and the hybrid window schemes improves persistence.
- KDLOR reaches the best results among all predictors. The way the thresholds are set gives the same importance to all classes, independently on their number of patterns.

Ordinal regression for the analysis of convective situations

D. Guijo-Rubio, C. Casanova-Mateo, J. Sanz-Justo, P.A. Gutiérrez, S. Cornejo-Bueno, C. Hervás-Martínez y S. Salcedo-Sanz. "Ordinal regression algorithms for the analysis of convective situations over Madrid-Barajas airport", Atmospheric Research, Vol. 236, 2020, pp. 104798.
JCR (2019): 4.676 Position: 13/93 (Q1).

Problem

- Meteorological events are associated with strong winds and local precipitations, which affects air and land operations at airports.
- Input variables data are obtained from a radiosonde station and from numerical weather models.
- The objective variable (convective clouds presence at the airport) is highly imbalanced.

Under-/Over-sampling + Dataset

Undersampling

- 30% of the training patterns labelled with CLEAR were randomly removed.

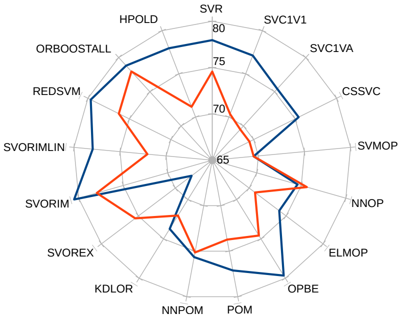
Class of day (C_q)	Training	Test
CLEAR	240 (68.18%)	82 (69.49%)
TCU	58 (16.48%)	20 (16.95%)
CB	28 (7.96%)	9 (7.63%)
TS	26 (7.39%)	7 (5.93%)
Total	352 (74.89%)	118 (25.11%)

Oversampling

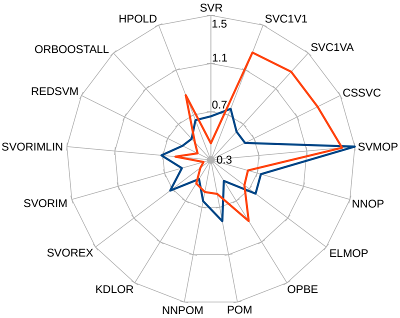
- Order among classes need to be preserved.
- OGO-ISP technique. Main goal is to create synthetic patterns only in the region within the minority classes (CB and TS).

Results

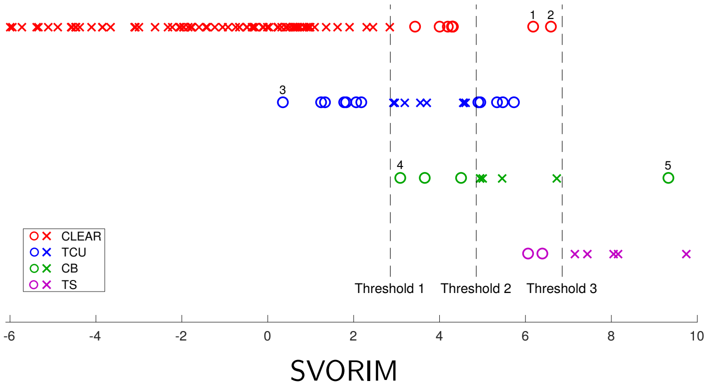
CCR



AMAE



Results



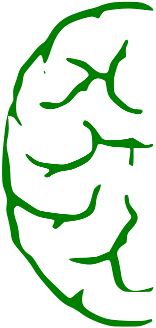
- Sudden changes in the previous and following day.
- Long runs of same sort of event.
- Distribution of the hourly labelling of this day.

Conclusions

Forecast source	SVORIM	TAF
CLEAR		
Hit Rate	0.90	0.80
False Alarm Rate	0.19	0.06
True Skill Score	0.71	0.75
TCU		
Hit Rate	0.40	0.90
False Alarm Rate	0.09	0.19
True Skill Score	0.31	0.71
CB		
Hit Rate	0.56	0.44
False Alarm Rate	0.08	0.03
True Skill Score	0.47	0.42
TS		
Hit Rate	0.71	0.71
False Alarm Rate	0.01	0.01
True Skill Score	0.71	0.71

- For CLEAR events, methods are similar. Good ability to separate "yes" events from "no" events.
- SVORIM is more skilful in predicting hits, whereas TAF is more skillful in avoiding false alarms.
- TCU (Cumulus Congestus) is easy for TAF given that is an intermediate stage between CLEAR and CB (Cumulonimbus).

Outline



1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Introduction

Time series prediction can be tackled from a more traditional point of view. In this sense, we have included three different perspectives:

- Nominal classification based on a multi-objective paradigm.
- Standard regression using advanced ANNs models.
- Multi-task models for regression.

Introduction

Time series prediction can be tackled from a more traditional point of view. In this sense, we have included three different perspectives:

- Nominal classification based on a multi-objective paradigm.
- Standard regression using advanced ANNs models.
- Multi-task models for regression.

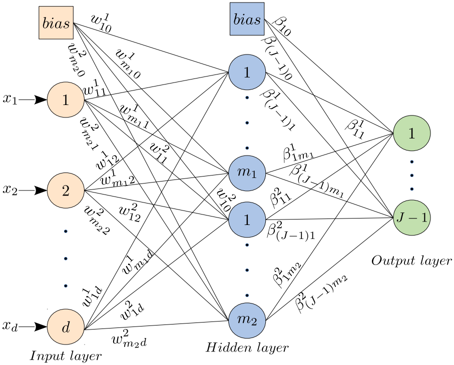
Prediction of convective clouds via multi-objective evolutionary techniques

D. Guijo-Rubio, P.A. Gutiérrez, C. Casanova-Mateo, J.C. Fernández, A.M. Gómez-Orellana, P. Salvador-González, S. Salcedo-Sanz and C. Hervás-Martínez. "Prediction of convective clouds formation using evolutionary neural computation techniques", Neural Computing and Applications, Vol. 32, 2020, pp. 13917 – 13929. JCR (2019): 4.774 Position: 23/136 (Q1).

Problem

- Very important problem in different areas such as agriculture, natural hazards prevention or transport-related facilities.
- The objective variable (convective clouds presence at the airport) is highly imbalanced.

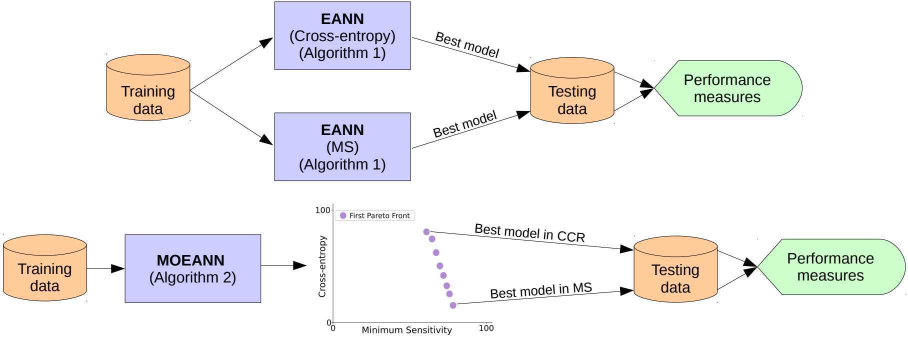
Methodology



$$f_q(\mathbf{x}, \mathbf{w}, \boldsymbol{\beta}) = \beta_{q0} + \sum_{j=1}^{m} \beta_{qj} B_j(\mathbf{x}, \mathbf{w}_j),$$
$$q = 1, \ldots, J-1, \quad (7)$$

$$f_q(\mathbf{x}, \theta) = \beta_{q0} + \sum_{j=1}^{m_1} \beta^1_{qj} B^1_j(\mathbf{x}, \mathbf{w}^1_j)$$
$$+ \sum_{j=1}^{m_2} \beta^2_{qj} B^2_j(\mathbf{x}, \mathbf{w}^2_j),$$
$$q = 1, 2, \ldots, J-1, \quad (8)$$

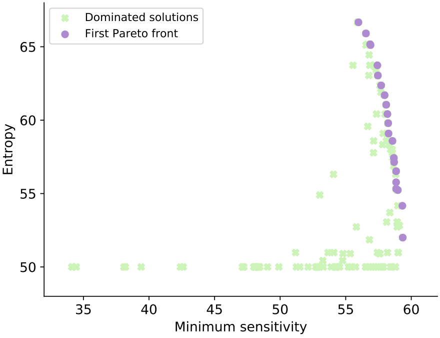
Methodology



Results

	Name	CCR	MS	AUC	#links
EANN-CCR	PU	72.062 ± 1.759	0.847 ± 3.249	84.396 ± 6.727	**42.900 ± 10.899**
	SU	**73.870 ± 1.667**	4.233 ± 6.159	88.300 ± 1.876	82.033 ± 13.828
	RBF	72.062 ± 1.612	0.370 ± 2.029	*88.857 ± 0.937*	76.633 ± 12.979
	PURBF	72.147 ± 1.193	0.000 ± 0.000	87.023 ± 2.774	69.900 ± 14.079
	SURBF	*73.503 ± 1.723*	0.476 ± 2.608	**88.938 ± 1.144**	75.700 ± 13.473
EANN-MS	PU	44.463 ± 5.849	*19.262 ± 10.637*	62.659 ± 9.057	52.467 ± 9.537
	SU	40.734 ± 7.687	15.405 ± 8.646	60.940 ± 7.417	67.000 ± 17.834
	RBF	42.034 ± 7.085	17.438 ± 12.886	65.142 ± 7.859	*46.600 ± 15.843*
	PURBF	41.610 ± 5.282	16.068 ± 10.708	62.051 ± 6.524	66.733 ± 18.260
	SURBF	40.876 ± 7.152	17.217 ± 10.496	61.919 ± 6.633	58.633 ± 18.277
MOEANN-CCR	PU	70.056 ± 1.051	0.166 ± 0.912	70.436 ± 9.983	59.200 ± 11.868
	SU	72.062 ± 1.787	0.333 ± 1.826	82.886 ± 4.243	83.800 ± 9.513
	RBF	71.638 ± 1.691	0.704 ± 2.339	86.944 ± 1.299	68.966 ± 17.058
	PURBF	71.780 ± 1.638	0.370 ± 2.029	85.896 ± 3.195	69.066 ± 11.316
	SURBF	72.514 ± 1.748	1.347 ± 3.705	86.599 ± 2.103	78.266 ± 10.913
MOEANN-MS	PU	40.508 ± 13.503	15.753 ± 11.003	61.433 ± 9.718	55.366 ± 12.081
	SU	50.367 ± 10.217	18.272 ± 10.790	70.462 ± 7.880	85.266 ± 8.808
	RBF	59.463 ± 7.850	13.913 ± 9.726	75.928 ± 6.305	70.933 ± 22.190
	PURBF	53.644 ± 11.906	16.958 ± 11.477	71.117 ± 9.321	72.066 ± 17.071
	SURBF	56.780 ± 8.999	**20.508 ± 11.072**	74.931 ± 5.020	77.133 ± 13.302

Conclusions



$$E = \frac{100}{1 + E_{min}} \quad (9)$$

Conclusions

- The Pareto front is wide, being a sign that it is correctly covering many different solutions.
- Mixing basis functions SU + RBF, along with a multi-objective strategy achieves the best results in terms of MS.

Introduction

Time series prediction can be tackled from a more traditional point of view. In this sense, we have included three different perspectives:

- Nominal classification based on a multi-objective paradigm.
- Standard regression using advanced ANNs models.
- Multi-task models for regression.

Evolutionary artificial neural networks for accurate solar radiation prediction

D. Guijo-Rubio, A.M. Durán-Rosal, P.A. Gutiérrez, A.M. Gómez-Orellana, C. Casanova-Mateo, J. Sanz-Justo, S. Salcedo-Sanz y C. Hervás-Martínez. "Evolutionary artificial neural networks for accurate solar radiation prediction", Energy, Vol. 210, 2020, pp. $1-11$. JCR (2019): 6.082 Position: 3/61 (Q1)

Problem

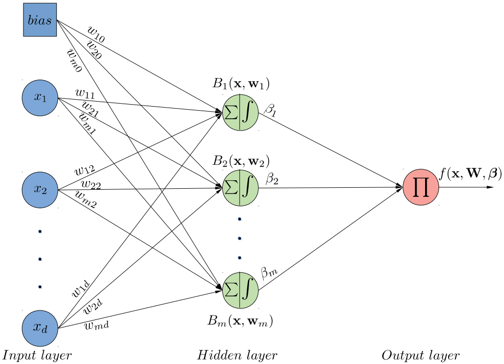
- Renewable energies such as the solar radiation have an inherent intermittence. Therefore, managing it, improves the solar energy availability and management.
- Solar radiation problem using only satellite-based measurements, avoiding the use of data from ground stations or atmospheric soundings (much more expensive and complicated).

Data

Predictive variables	units
Reflectivity (VIS 0.6 and VIS 0.8 channels)	[%]
Clear sky radiance	[W/m^2]
Cloud index	[%]
CAMS solar radiation	[W/m^2]
SolarGIS solar radiation	[W/m^2]
Target	**units**
Global solar radiation	[W/m^2]

- Use of Heliosat-2, CAMS and SolarGIS numerical models.
- 4 different configurations, ranging from 5 input variables to 40.

Methodology



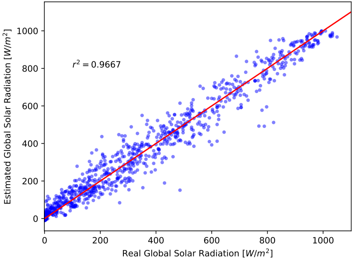
$$B_j(\mathbf{x}, \mathbf{w}_j) = \frac{1}{1 + e^{-\left(w_{j0} + \sum_{i=1}^{d} w_{ji} x_i\right)}}, \quad j = 1, \ldots, m. \quad (10)$$

$$f(\mathbf{x}, \mathbf{W}, \boldsymbol{\beta}) = \prod_{j=1}^{m} B_j(\mathbf{x}, \mathbf{w}_j)^{\beta_j}. \quad (11)$$

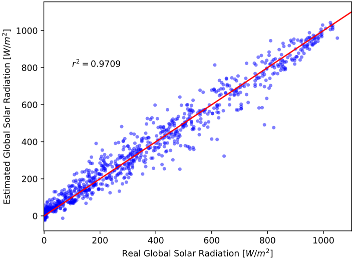
Results

	Best RMSE model				
	MBE $[W/m^2]$	MAE $[W/m^2]$	RMSE $[W/m^2]$	r^2	$\overline{\mathrm{RMSE}}$ $[W/m^2]$
Configuration 1					
SU-LO	2.26	39.17	59.06	0.9622	60.73 ± 0.85
RBF-LO	2.70	37.33	57.07	0.9648	62.31 ± 4.79
SU-PU	1.58	38.64	58.41	0.9630	59.99 ± 0.90
Configuration 2					
SU-LO	1.92	34.59	*51.89*	*0.9708*	$\mathit{55.29 \pm 1.52}$
RBF-LO	1.64	36.27	55.47	0.9667	61.54 ± 6.85
SU-PU	1.09	**33.46**	**51.82**	**0.9709**	57.02 ± 9.27
Configuration 3					
SU-LO	2.55	33.88	52.46	0.9702	$\mathbf{55.09 \pm 2.79}$
RBF-LO	1.42	34.11	52.74	0.9698	63.05 ± 22.66
SU-PU	1.70	*33.49*	52.17	0.9705	57.63 ± 10.07
Configuration 4					
SU-LO	-2.95	36.95	56.15	0.9659	67.42 ± 13.92
RBF-LO	$\mathit{-1.04}$	36.92	55.14	0.9670	60.38 ± 4.85
SU-PU	**0.30**	33.61	52.12	0.9705	56.38 ± 2.67

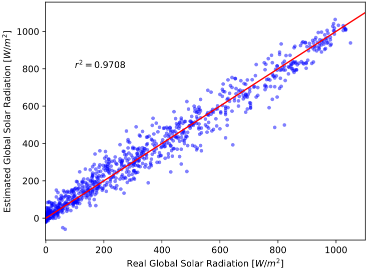
Results



RBF-LO



SU-PU



SU-LO

Conclusions

Conclusions: configurations

- All the proposed configurations reach to good results, making the proposed methodology robust.
- The best configuration includes 8 input variables, including Heliosat-2, CAMS and SolarGIS models.

Conclusions: solar radiation

- The proposed EANNs are able to obtain an extremely accurate prediction of solar radiation, only based on satellite measurements.
- The proposed approach can be seen as a post-processing step to the CAMS and SolarGIS methods.

Introduction

Time series prediction can be tackled from a more traditional point of view. In this sense, we have included three different perspectives:

- Nominal classification based on a multi-objective paradigm.
- Standard regression using advanced ANNs models.
- Multi-task models for regression.

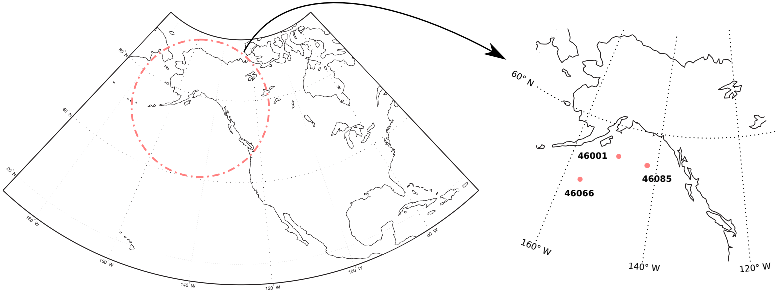
Energy flux prediction using multi-task EANNs

D. Guijo-Rubio, A.M. Gómez-Orellana, P.A. Gutiérrez y C. Hervás-Martínez. "Short- and long-term energy flux prediction using Multi-Task Evolutionary Artificial Neural Networks", Ocean Engineering, Vol. 216, 2020, pp.108089. JCR (2019): 3.068 Position: 1/14 (Q1).

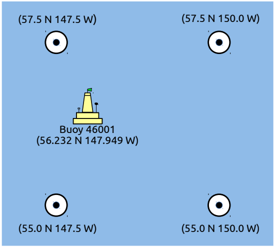
Problem

- Waves exhibit a stochastic nature, due to the influence of a great number of environmental elements. Therefore, they cannot be predicted straightforwardly.
- Useful for energy storage systems and a cost-effective transmission of electricity.

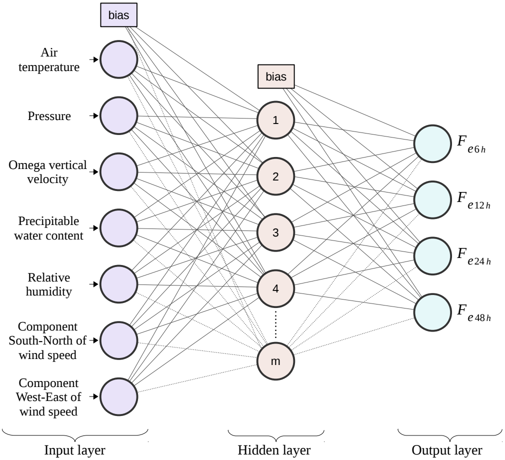
Data



Variable	Units
Air temperature	degK
Pressure	Pascals
Omega vertical velocity	Pascal/s
Precipitable water content	kg/m^2
Relative humidity	%
Component South-North of wind speed	m/s
Component West-East of wind speed	m/s



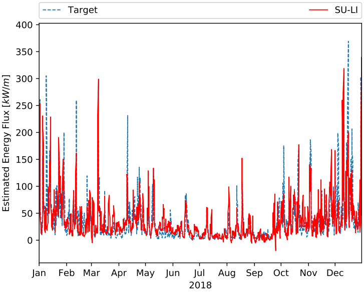
Methodology

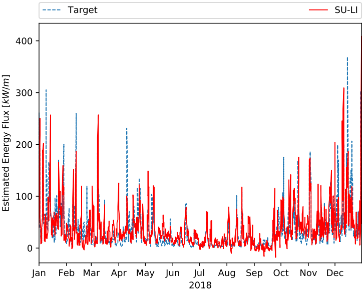


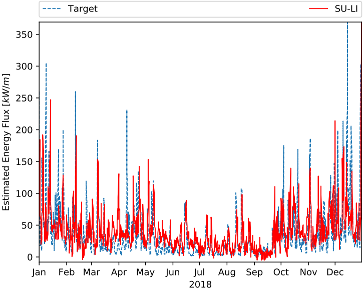
Methodology

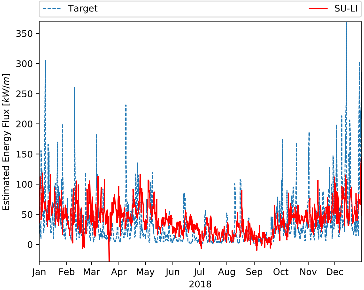
- Use of an EA as an efficient search method.
- Use of SU, PU and RBFs in the hidden layer.
- Multi-task learning.

Results


Target
SU-LI
Estimated Energy Flux [kW/m]
400
350
300
250
200
150
100
50
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2018


Target
SU-LI
Estimated Energy Flux [kW/m]
400
300
200
100
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2018


Target
SU-LI
Estimated Energy Flux [kW/m]
350
300
250
200
150
100
50
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2018


Target
SU-LI
Estimated Energy Flux [kW/m]
350
300
250
200
150
100
50
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2018

Conclusions

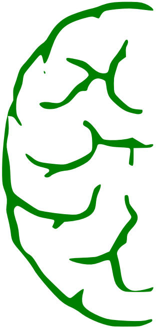
Conclusions: energy flux

- The prediction of the energy flux is performed only considering reanalysis data, avoiding missing data problems and allowing the applicability to other locations.
- Anticipating, not only to short-term phenomena (6 hours), but also long-term (2 days).

Conclusions: methodology

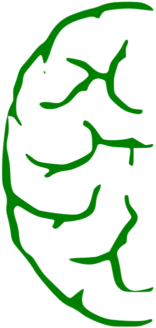
- Goal: obtain a multi-task model (4 time prediction horizons) achieving competitive results with lower-complexity.
- SU-LI outperforms the remaining techniques.
- SU-PU achieves second best results, but shows less stability in the average results.

Outline



1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Outline



1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Hybrid approach to time series classification with shapelets

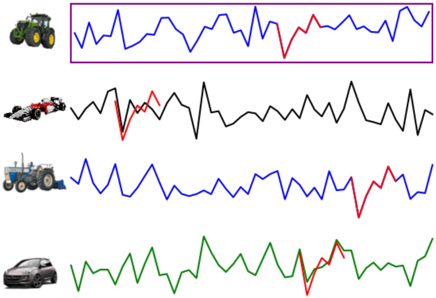
D. Guijo-Rubio, P.A. Gutiérrez, R. Tavenard y A. Bagnall. “A hybrid approach to time series classification with shapelets”. 20th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2019). 2019. LNCS, Vol. 11871, pp. 137 – 144.

Problem

- Is there any approach reaching the state-of-the-art performance with the lowest time complexity?
- Create a better classifier by hybridising data-driven search and stochastic gradient descent learning.

Two research stays in the University of East Anglia, and in collaboration with the Alan Turing Institute in `sktime`.

Methodology



Shapelet extraction framework

1. Candidate generation (with a length constraint)
2. Measuring similarity between the candidate and the time series (minimum).
3. Measuring quality of the candidate.

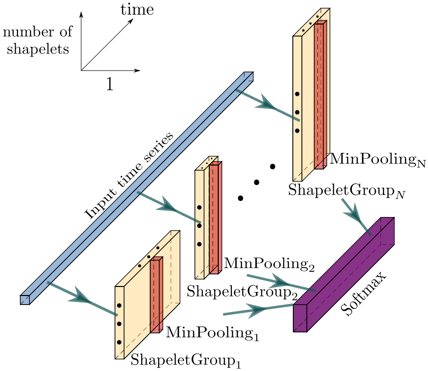
Shapelet transform

1. Only good shapelets are retained.
2. Attributes are the distance between shapelets and time series.
3. Any standard classifier could be applied to the transformed dataset.

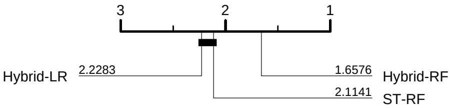
Methodology

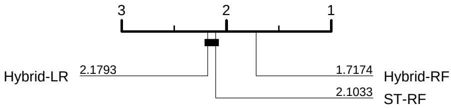
Steps

1. Sample shapelets for a fixed time (contract shapelets).
2. Tune these shapelets with the learning shapelets algorithm.



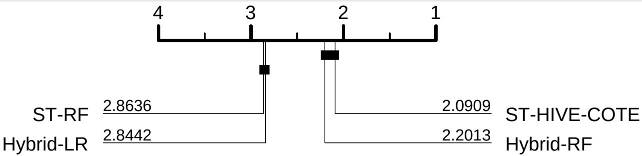
Results and conclusions



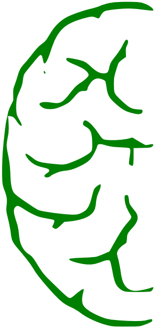


Conclusions

1. Tuning significantly improved accuracy after a 1 hour and 10 hours search.
2. Similar results to the state-of-the-art approach are obtained.



Outline



1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Time series ordinal classification

D. Guijo-Rubio, P.A. Gutiérrez, A. Bagnall y C. Hervás-Martínez. "Time series ordinal classification via shapelets". 2020 IEEE International Joint Conference on Neural Networks (IJCNN 2020). 2020. Glasgow, UK. pp. $1-8$.

D. Guijo-Rubio, P.A. Gutiérrez, A. Bagnall y C. Hervás-Martínez. "Ordinal versus nominal time series classification". 5th Workshop on Advances Analytics and Learning on Temporal Data (AALTD 2020). 2020. LNAI, Vol. 12588, pp. $19-29$.

Problem

- No specific ordinal approaches in the literature. This Thesis proposed some techniques for this novel field.
- Demonstrate that, for ordinal datasets, the ordinal approaches are able to achieve better performance than standard TSC techniques, in terms of accuracy.

Methodology

Framework of the ST → ordinal information in two points:

1. Shapelet extraction procedure:
   1. Candidate generation, i.e. generation of a subsequence satisfying the previous length constraint.
   2. Measuring similarity between the candidate and the time series.
   3. Measuring quality of the candidate.
2. Final classifier: ordinal techniques.

Quality of the candidate:

$$OF(\mathbf{s}) = \frac{\sum_{k=1}^{Q} \sum_{j=1}^{Q} |k-j|(\bar{x}_k - \bar{x}_j)^2}{(Q-1)\sum_{k=1}^{Q}(S_k)^2} \quad (12)$$

$$R^2(\mathbf{s}) = \frac{S(d_{\mathbf{s},\mathbf{T}_i}, c_{\mathbf{s},\mathbf{T}_i})}{S_{d_{\mathbf{s},\mathbf{T}_i}} S_{c_{\mathbf{s},\mathbf{T}_i}}} \quad (13)$$

$$\rho(\mathbf{s}) = 1 - \frac{6\sum_{i=1}^{N} D(\mathbf{s}, \mathbf{T}_i)^2}{N(N^2-1)} \quad (14)$$

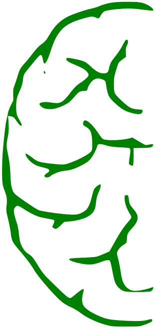
Ordinal versus nominal time series classification

	SVC1V1	SVC1VA	SVORIM	HIVE-COTE	InceptionTime	TS-CHIEF
DistalPhalanxOutline	*74.82*	74.10	**75.54**	**75.54**	73.38	74.10
DistalPhalanxTW	**70.50**	69.06	*69.78*	67.63	68.35	68.35
EthanolLevel	61.00	58.80	62.40	*71.40*	**81.40**	52.80
MiddlePhalanxOutline	*62.34*	**63.64**	**63.64**	59.09	55.19	59.09
MiddlePhalanxTW	**59.09**	*56.49*	*56.49*	55.84	51.30	55.85
ProximalPhalanxOutline	85.85	*86.34*	**87.32**	84.39	84.88	84.88
ProximalPhalanxTW	72.68	76.59	76.10	*80.00*	77.56	**81.46**
Average ranking	*3.00*	3.21	**2.36**	3.79	4.43	4.21
#Wins	*2*	1	**3**	1	1	1

Conclusions

- ST combined with R^2 is able to achieve the best results (independently of the OC), specially in terms of *AMAE*.
- SVORIM + ST R^2 $\rightarrow$ obtains the best performance in terms of *CCR*.
- Nominal classifiers, SVC1V1 and SVC1VA, are taking advantage of the ordinal information, obtaining competitive results, better than those of the ensemble approaches.

Outline



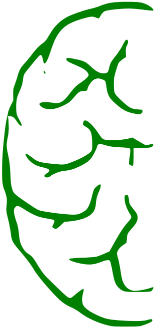
1 Introduction
2 Background
3 Objectives
4 Time series clustering
5 Time series prediction
- Time series ordinal prediction
- Time series forecasting

6 Time series classification
- Time series classification
- Time series ordinal classification

7 Non-temporal data mining
- Non-temporal data regression
- Non-temporal data classification

8 Conclusions and future works

Outline



1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Modelling of engineering applications

F. Comino, **D. Guijo-Rubio**, M. R. de Adana and C. Hervás-Martínez. "Validation of multitask artificial neural networks to model desiccant wheels activated at low temperature", International Journal of Refrigeration, Vol. 100. 2019, pp. 434 – 442. JCR (2019): 3.461 Position: 11/61 (Q1).

F.J. Jiménez-Romero, **D. Guijo-Rubio**, F.R. Lara-Raya, A. Ruiz-González y C. Hervás-Martínez. "Validation of artificial neural networks to model the acoustic behaviour of induction motors", Applied Acoustics, Vol. 166, 2020, pp. 107332. JCR (2019): 2.440 Position: 9/32 (Q2).

Outline

1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Health-based problems

A. Rivero-Juárez, **D. Guijo-Rubio**, F. Téllez, R. Palacios, D. Merino, J. Macías, J.C. Fernández, P.A. Gutiérrez, A. Rivero and C. Hervás-Martínez. "Using machine learning methods to determine a typology of patients with HIV-HCV infection to be treated with antivirals", PLoS One, Vol. 15(1). 2020, pp. e0227188. JCR (2019): 2.740 Position: 27/71 (Q2).

D. Guijo-Rubio, P.J. Villalón-Vaquero, P.A. Gutiérrez, M.D. Ayllón, J. Briceño y C. Hervás-Martínez. "Modelling survival by machine learning methods in liver transplantation: application to the UNOS dataset". 20th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2019). 2019. LNCS, Vol. 11872, pp. 97 – 104.

D. Guijo-Rubio, J. Briceño, P. A. Gutiérrez, M.D. Ayllón, R. Ciria, C. Hervás Martínez. "Comparison of statistical methods and machine learning techniques for donor-recipient matching in liver transplantation". PLos One, 2021. JCR (2019): 2.740 Position: 27/71 (Q2).

Outline

1. Introduction
2. Background
3. Objectives
4. Time series clustering
5. Time series prediction
   - Time series ordinal prediction
   - Time series forecasting
6. Time series classification
   - Time series classification
   - Time series ordinal classification
7. Non-temporal data mining
   - Non-temporal data regression
   - Non-temporal data classification
8. Conclusions and future works

Conclusions I - Clustering

- Novel approach to time series clustering, consisting in grouping time series based on their similarity.
- First stage: growing window segmentation + segments projection into a fixed-size vector of statistical characteristics + clustering new segment representation.
- Second stage: a common structure for the time series is built by including information of the centroids and of the segments with the highest variance + hierarchical clustering, grouping this novel time series representation by their similarity.

This block satisfies objectives 3 and 4.

Conclusions II - Prediction

As an OC task

- Ordinal nature among the labels → use of ordinal classifiers for taking advantage of the ordinal information.
- We proposed the use of three different windows based on AR models and ordinal oversampling methods for balancing the datasets.
- Comprehensive comparison against other ordinal classifiers and against physical models.

As traditional ML tasks

- Different ML paradigms:
    - Multi-objective point of view (CCR vs MS).
    - Novel mixture for ANNs: SUs in hidden layer with PUs in the output layer.
    - Multi-task learning applied to EANNs.

This block partially satisfies objectives 1, 2, 3, and 7.

Conclusions III - Classification

TSC

- Development of a hybrid model $\rightarrow$ standard ST and LS approaches.
- The results achieved by this hybrid method are significantly better than either approach in isolation.
- The results achieved are similar to SOTA, being less computationally intensive.

TSOC

- Three different shapelet quality measures considering ordinal information, based on adaptations to the ordinal paradigm of traditional indices + use of ordinal classifiers applied to the transform.
- HIVE-COTE, TS-CHIEF and inceptionTime are outperformed by the ST version adapted to OC.

This block partially satisfies objectives 3, 5, and 6.

Conclusions IV - Non-temporal data

Engineering applications

- These engineering applications typically concern more than one objective, trying to optimise them all simultaneously.
- Modelling of desiccant wheels and acoustic behaviour of induction motors.

Health-based problems

- We tackled the HIV/HCV disease (typology of co-infected patient to be treated).
- The LT problem has been tackled from two points of view: survival analysis and donor-recipient matching.

This block partially satisfies objectives 1, 2, and 7.

Future lines

- The use of optimal time series segmentation techniques could improve the performance of the time series clustering technique.

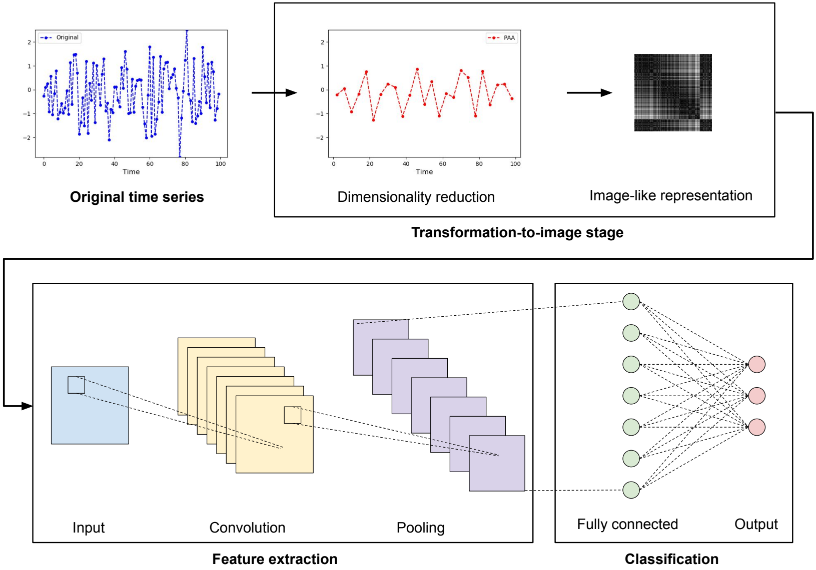
Future lines

- The use of optimal time series segmentation techniques could improve the performance of the time series clustering technique.
- The prediction of convective situations can be solved by means of label distribution learning techniques. It consists in predicting the degree to which each label describes the instance.

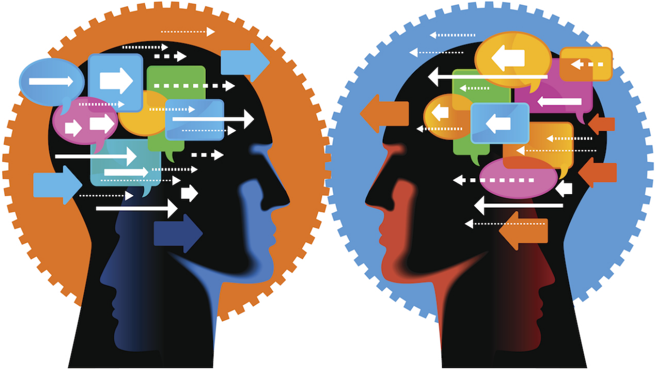
Future lines

- The use of optimal time series segmentation techniques could improve the performance of the time series clustering technique.
- The prediction of convective situations can be solved by means of label distribution learning techniques. It consists in predicting the degree to which each label describes the instance.
- Transforming 1D time series to 2D image-like representation is a recent research line being tackled at the moment of writing this Thesis.

1D time series to 2D image-like representation



Thank you for your attention!



Questions?

Clustering, prediction and ordinal classification of time series using machine learning techniques: applications

International PhD

David Guijo Rubio
dguijo@uco.es
Supervisors:
César Hervás Martínez
Pedro Antonio Gutiérrez Peña

Dept. of Computer Science and Numerical Analysis, University of Córdoba.
Learning and Artificial Neural Networks (AYRNA) research group.

21st June 2021

UNIVERSIDAD
D
CÓRDOBA

Methodology

Time series clustering:

```
Require: Time series dataset
Ensure: Best quality clustering
for Each time series do
    Apply time series segmentation
    for Each segment do
        Extract the coefficients of the segment
        Compute the statistical features
        Combine the coefficients and the statistical features into a single array
    end for
    Cluster all the mapped segments
    Based on the previous clustering, map each time series
end for
Cluster mapped time series
Evaluate the goodness of the clustering
return Best quality clustering
```

Segments mapping

Segments mapping: each segment is projected into:

$$\mathbf{v}_s = (\mathbf{p}_s, S_s^2, \gamma_{1_s}, AC_s)$$

$$S_s^2 = \frac{1}{t_s - t_{s-1} + 1} \sum_{i=t_{s-1}}^{t_s} (y_i - \overline{y_s})^2 \tag{15}$$

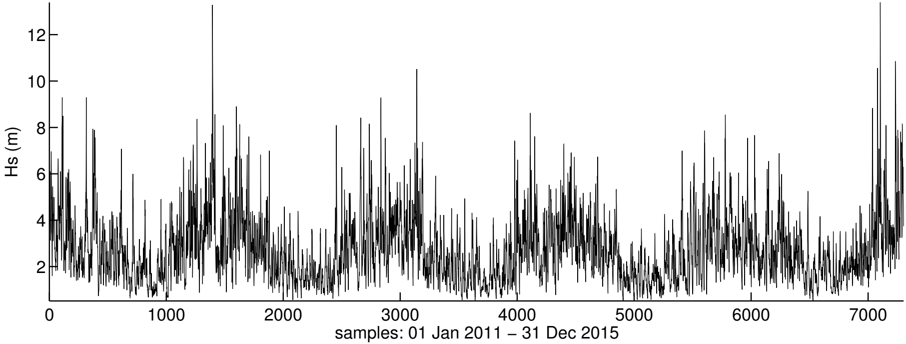
$$\gamma_{1s} = \frac{\frac{1}{t_s - t_{s-1}+1} \sum_{i=t_{s-1}}^{t_s} (y_i - \overline{y_s})^3}{\hat{\sigma}_s^3} \tag{16}$$

$$AC_s = \frac{\sum_{i=t_{s-1}}^{t_s} (y_i - \overline{y_s}) \cdot (y_{i+1} - \overline{y_s})}{S_s^2} \tag{17}$$

Statistical features associated to moments higher than three (e.g. kurtosis) are not considered, because the segments obtained are usually short, and they are not able to provide additional relevant information.

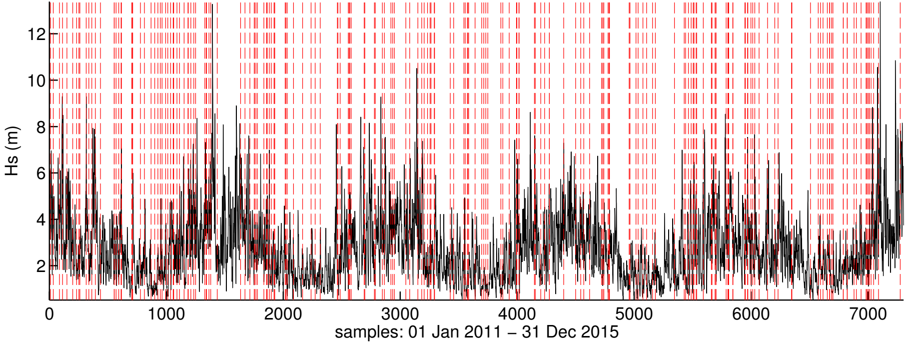
Time series segmentation

Given a time-series $\mathbf{Y} = \{y_n\}_{n=1}^{N}$.



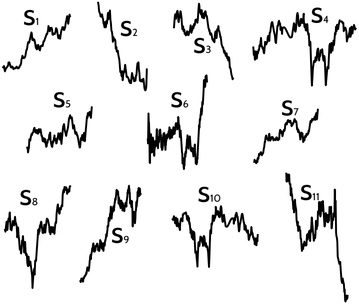
Time series segmentation

Find m segments, defined by: $t_1 < t_2 < t_{m-1}$.



Time series segmentation

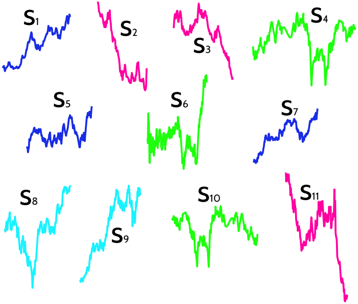
$$S_1 = \{y_1, \ldots, y_{t_1}\}, S_2 = \{y_{t_1}, \ldots, y_{t_2}\}, \ldots, S_m = \{y_{t_{m-1}}, \ldots, y_N\}$$



Extract the segments, and then we have two objectives.

Time series segmentation

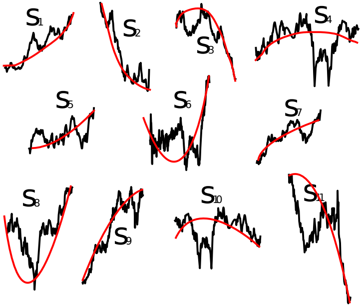
$$S_1 = \{y_1, \ldots, y_{t_1}\}, S_2 = \{y_{t_1}, \ldots, y_{t_2}\}, \ldots, S_m = \{y_{t_{m-1}}, \ldots, y_N\}$$



Associate a label (colour) to each segment.

Time series segmentation

$$S_1 = \{y_1, \ldots, y_{t_1}\}, S_2 = \{y_{t_1}, \ldots, y_{t_2}\}, \ldots, S_m = \{y_{t_{m}-1}, \ldots, y_N\}$$



Or to approximate each segment with a regression or interpolation.

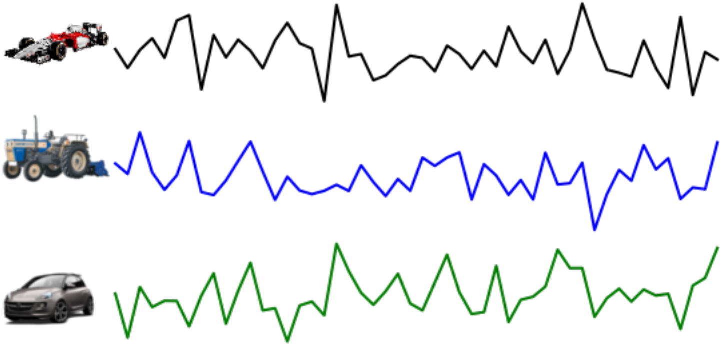
Shapelets

Definition

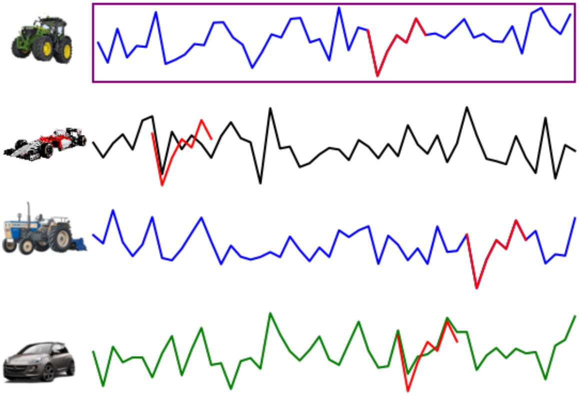
Discriminatory phase independent subsequences forming a basic primitive for TSC algorithms.

Shapelet: metal fender and hood.

Shapelets

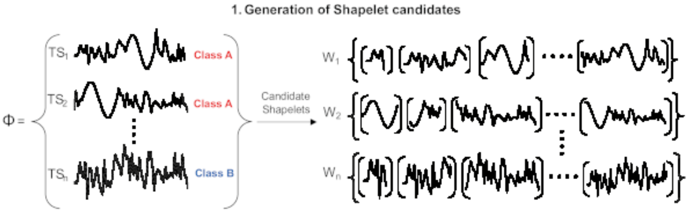


Shapelets

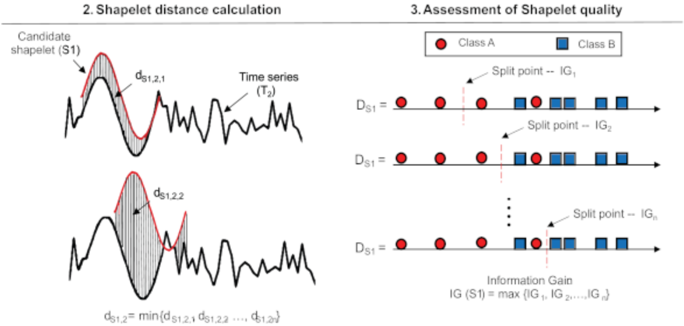


Effective tool for TSC and popular research topic.

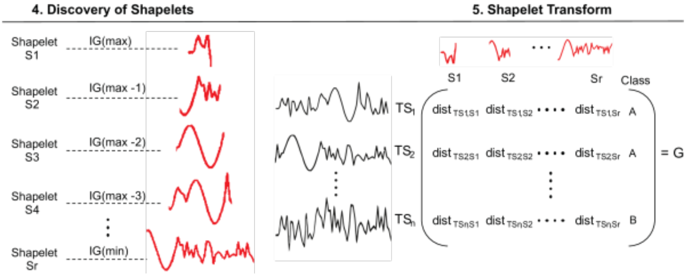
Shapelets


1. Generation of Shapelet candidates
Φ =
TS1
Class A
TS2
Class A
TSn
Class B
Candidate Shapelets
W1
W2
Wn

Shapelets

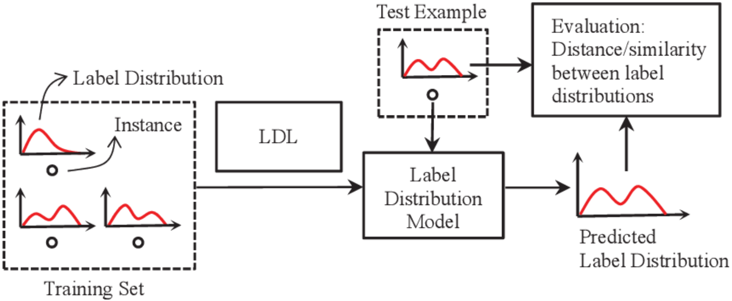

2. Shapelet distance calculation
3. Assessment of Shapelet quality
Candidate shapelet (S1)
$d_{S1,2,1}$
Time series (T_2)
$d_{S1,2,2}$
$d_{S1,2} = \min\{d_{S1,2,1}, d_{S1,2,2}, \ldots, d_{S1,2n}\}$
Class A
Class B
Split point -- IG_1
Split point -- IG_2
Split point -- IG_n
D_{S1} =
Information Gain
$IG(S1) = \max\{IG_1, IG_2, \ldots, IG_n\}$

Shapelets

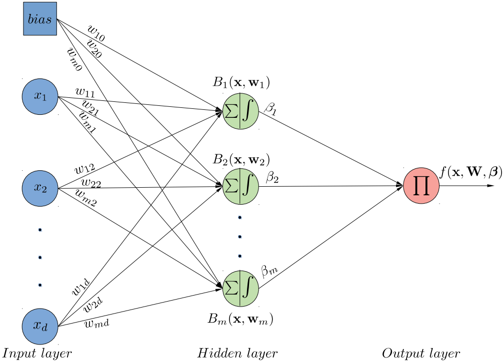

4. Discovery of Shapelets
5. Shapelet Transform
Shapelet S1
IG(max)
Shapelet S2
IG(max -1)
Shapelet S3
IG(max -2)
Shapelet S4
IG(max -3)
Shapelet Sr
IG(min)
S1
S2
Sr
Class
TS_1
TS_2
TS_n
$dist_{TS1S1}$ $dist_{TS1S2}$ $dist_{TS1Sr}$ A
$dist_{TS2S1}$ $dist_{TS2S2}$ $dist_{TS2Sr}$ A
$dist_{TSnS1}$ $dist_{TSnS2}$ $dist_{TSnSr}$ B
= G

Label Distribution Learning (LDL)

(a) Single-label learning (b) Multi-label learning (c) Label distribution learning



Combination of activation functions



$$B_j(\mathbf{x}, \mathbf{w}_j) = \frac{1}{1 + e^{-\left(w_{j0} + \sum_{i=1}^{d} w_{ji} x_i\right)}}, \quad j = 1, \ldots, m. \quad (18)$$

$$f(\mathbf{x}, \mathbf{W}, \boldsymbol{\beta}) = \prod_{j=1}^{m} B_j(\mathbf{x}, \mathbf{w}_j)^{\beta_j}. \quad (19)$$

Combination of activation functions

The main reason behind this idea is to take advantage of the interactions between the outputs of the hidden layer, making the ANN more complex but accurate.

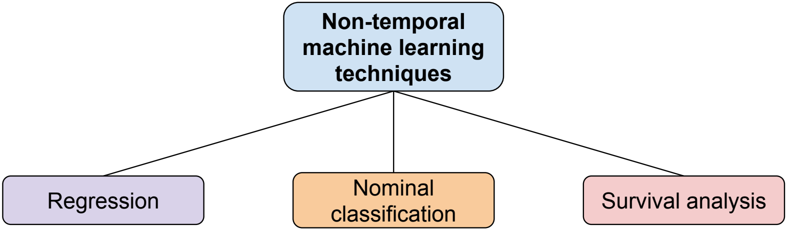
$$B_j(\mathbf{x}, \mathbf{w}_j) = \frac{1}{1 + e^{-\left(w_{j0} + \sum_{i=1}^{d} w_{ji} x_i\right)}}, \quad j = 1, \ldots, m. \tag{20}$$

Why SU and not RBF?

This mixture of functions is not adequate because RBFs are local functions, and evaluating their interaction would make no sense.

$$B_j(\mathbf{x}, \mathbf{w}_j) = e^{-\frac{1}{2}\left(\frac{\sum_{i=1}^{d}(x_i - c_{ji})^2}{r_j}\right)}, \quad j = 1, \ldots, m, \tag{21}$$

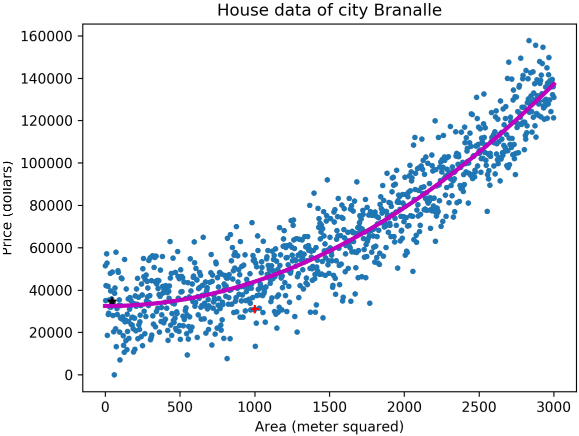
Non-temporal machine learning: **machine learning techniques**



Others: clustering, association, preprocessing, ...

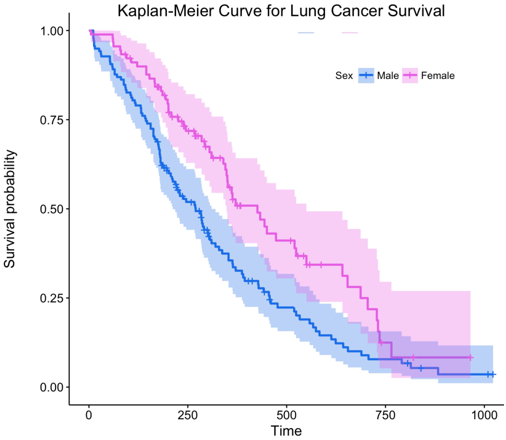
Non-temporal machine learning: **regression**

Given a pattern $\mathbf{x}_j$, the regression consists in estimating the value for a continuous output y_j.



Non-temporal machine learning: **survival analysis**

The survival analysis task consists in analysing the expected duration of time until the occurrence of an event.



Modelling of engineering applications

F. Comino, **D. Guijo-Rubio**, M. R. de Adana and C. Hervás-Martínez. "Validation of multitask artificial neural networks to model desiccant wheels activated at low temperature", International Journal of Refrigeration, Vol. 100. 2019, pp. 434 – 442. JCR (2019): 3.461 Position: 11/61 (Q1).

F.J. Jiménez-Romero, **D. Guijo-Rubio**, F.R. Lara-Raya, A. Ruiz-González y C. Hervás-Martínez. "Validation of artificial neural networks to model the acoustic behaviour of induction motors", Applied Acoustics, Vol. 166, 2020, pp. 107332. JCR (2019): 2.440 Position: 9/32 (Q2).

Problem

- The processes of these engineering applications entail considerable economic expenses. Therefore, modelling them enables their optimisation and allows their analysis.

Modelling of engineering applications

Conclusions

- Models obtained present low computational cost and good accuracy.
- Simplicity and interpretability are important characteristics of these models.
- Multitask ANN models have an effective transfer mechanism to extract common features from multiple tasks.

Determining a typology of patients with HIV/HCV infection

A. Rivero-Juárez, **D. Guijo-Rubio**, F. Téllez, R. Palacios, D. Merino, J. Macías, J.C. Fernández, P.A. Gutiérrez, A. Rivero and C. Hervás-Martínez. "Using machine learning methods to determine a typology of patients with HIV-HCV infection to be treated with antivirals", PLoS One, Vol. 15(1). 2020, pp. e0227188. JCR (2019): 2.740 Position: 27/71 (Q2).

Problem

- Identify those factors for HIV/HCV co-infected patients (to which clinicians have given careful consideration before treatment uptake) that have not being included among the prioritisation criteria.
- Find a simple model able allowing to analyse the relationship between patient characteristics and the probability of belonging to the treated group.

Determining a typology of patients with HIV/HCV infection

Conclusions

- The variable "Recent PWID" is mandatory. It represents the main limiting factor related to the absence of treatment uptake.
- The parsimony of the model is attractive since no extra useless information is needed from the patient and therefore minimises the likelihood of incurring in information errors.
- ANN models should be consider for drawing up or modifying strategic plans when tackling different diseases, given their potential as clinical decision making systems.

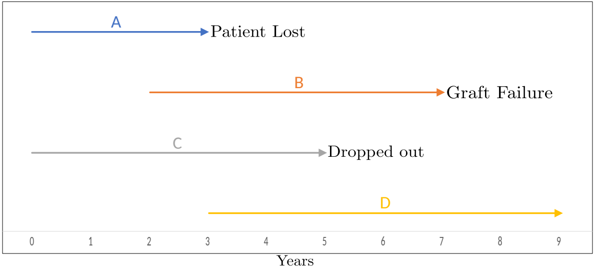
Modelling survival in liver transplantation

D. Guijo-Rubio, P.J. Villalón-Vaquero, P.A. Gutiérrez, M.D. Ayllón, J. Briceño y C. Hervás-Martínez. "Modelling survival by machine learning methods in liver transplantation: application to the UNOS dataset". 20th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2019). 2019. LNCS, Vol. 11872, pp. 97 – 104.

Problem

- Application of machine learning techniques for the survival analysis.
- Analysis of the performance on the largest database of liver transplant: UNOS database.

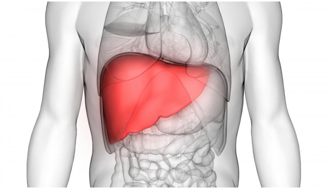
Methodology



Methods

- Cox's-regression-based models: Coxnet, CoxPH and IPCRidge.
- Models based on Gradient Boosting: GradientBoosting and ComponentwiseGB.
- Adaptations of SVM: FastSurvivalSVM and FastKernelSurvivalSVM.

Conclusions



Conclusions

- Survival analysis technique enable the use of censored data (74%).
- Evaluation of these models in the largest database provided by UNOS.
- Similar results obtained by all the techniques.

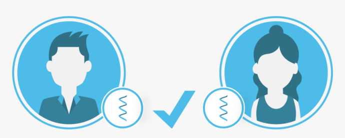
Donor-recipient matching in liver transplantation

D. Guijo-Rubio, J. Briceño, P. A. Gutiérrez, M.D. Ayllón, R. Ciria, C. Hervás Martínez. "Comparison of statistical methods and machine learning techniques for donor-recipient matching in liver transplantation". PLos One, 2021. JCR (2019): 2.740 Position: 27/71 (Q2).

Problem

- The increasing number of candidates for liver transplantation and the scarce number of available donors, the rationale for assignment of a given donor to potential candidates on a waiting list is a matter of controversy.
- Several scores have been designed, however, their main objective is to decrease the mortality in the waiting list without affecting the result of the transplant.

Methodology



- **Goal**: analyse how several machine learning techniques behave in the largest liver transplant database.
- **Methods**: statistical methods versus machine learning techniques.
- **Data**: UNOS database with more than 170.000 patients (donors+recipients).
- **Proposal**: Rule-based system trying to achieve a balance between graft survival and MELD.

Conclusions

- Logistic regression achieves the best performance. Improving popular scores such as MELD.
- Electronic Health Records have been developed to speed up the mechanism for clinician decision making. Not working:
  1. Missing values and the imputation techniques used.
  2. Increasing quantity of different categories for some attributes.
  3. Increasing number of "non-specified" cases in this attributes.
  4. Attributes with several categories but a small number of cases per category.
  5. Incongruities between different expert opinions.
- We provided the medical community with a tool bridging the gap between the medical decision (subjectivity) and strict mathematical scores (objectivity).