Aplicación de técnicas de selección de características para la mejora de los sistemas automáticos de detección de vertidos de hidrocarburos

David Mera [1] Veronica Bolon-Canedo [2]
J.M Cotos [1] Amparo Alonso-Betanzos [2]

[1] Centro Singular de Investigación en Tecnoloxías da Información
Universidade de Santiago de Compostela

[2] Dept. de Computación, Universidade da Coruña

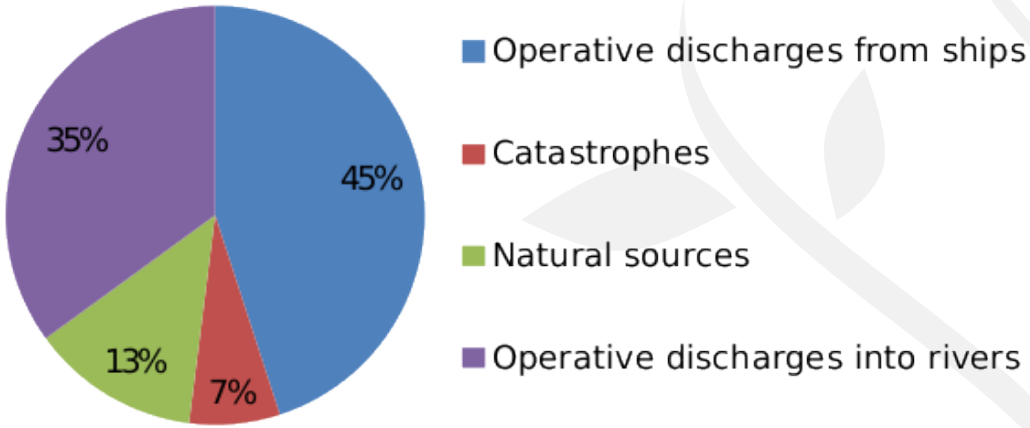
Machine Learning Workhop Galicia 2016

Index

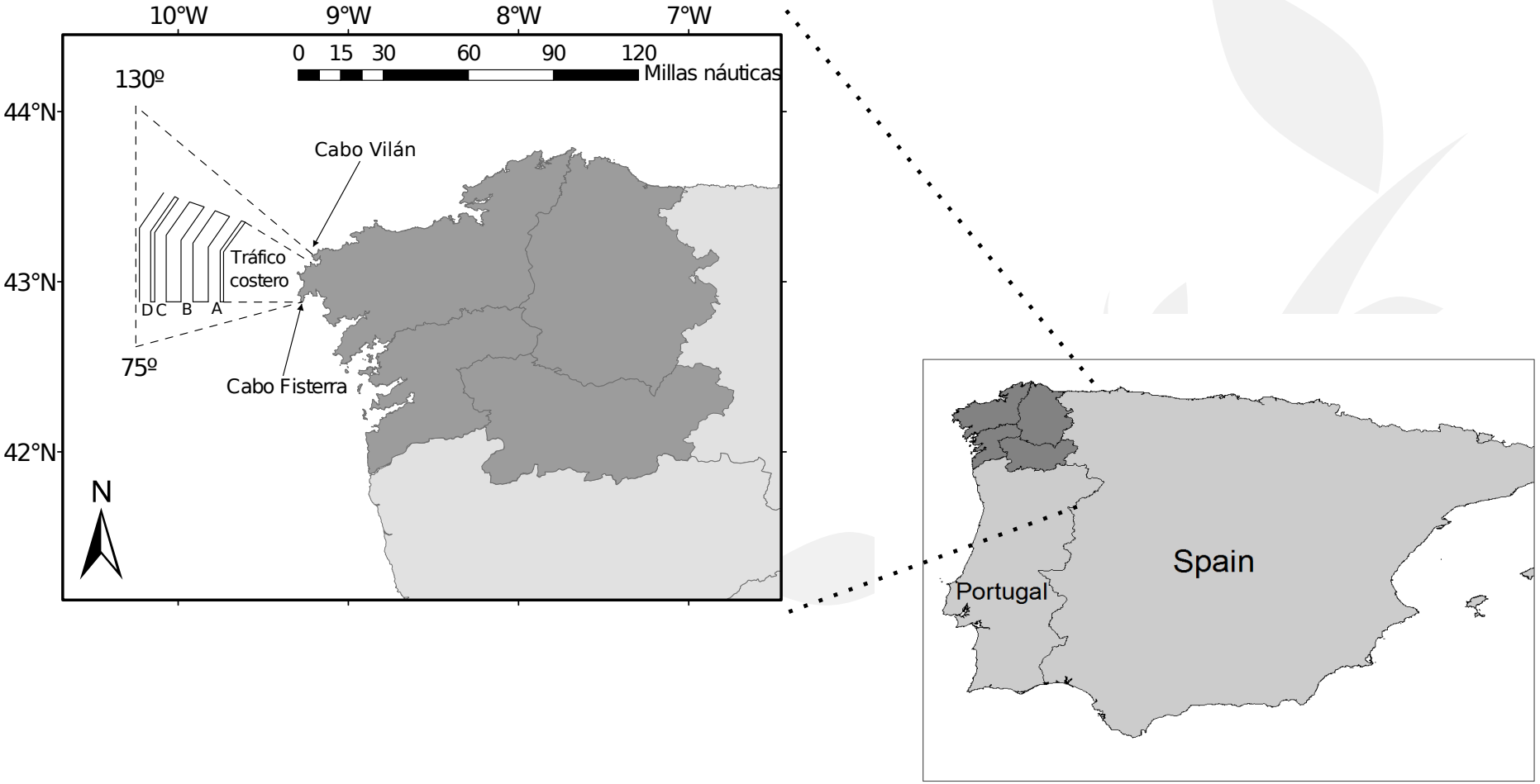
Index

Sentinazos Introduction

- The international trade is mainly supported by maritime transport
- The intensive traffic sails along the Exclusive Economic Zones (EEZ) of countries and generates important pollution problems
- Only the 7 % of oil spills come from catastrophes like tanker and oil platform accidents



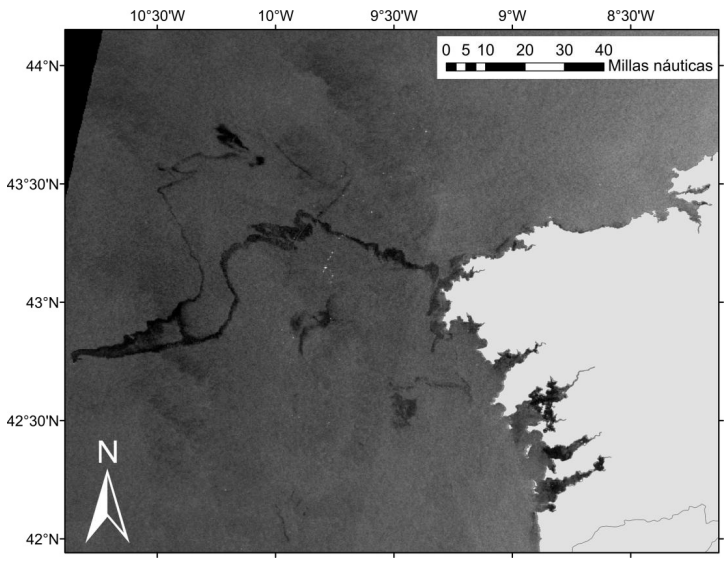
Sentinazos Introduction



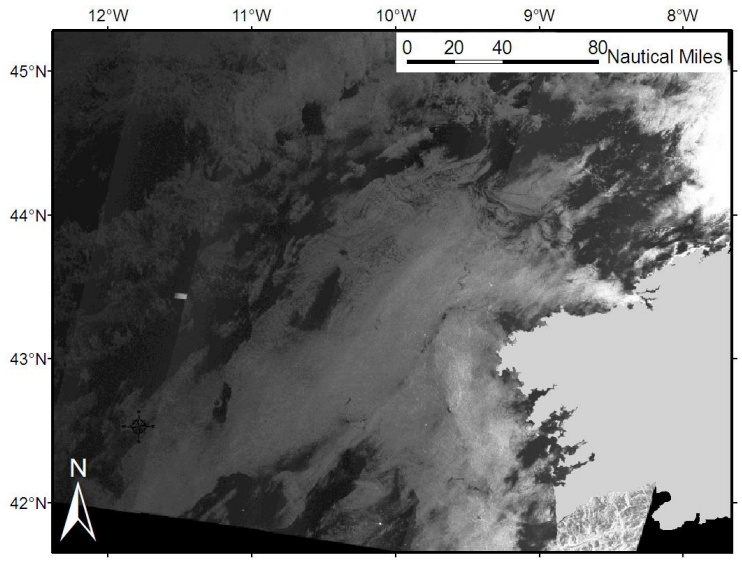
Sentinazos Introduction

Sentinazos Introduction

- An effective monitoring system is crucial to ensure a proper response to environmental emergencies
- The Synthetic Aperture Radar has proved to be an appropriate tool for discovering oil spills

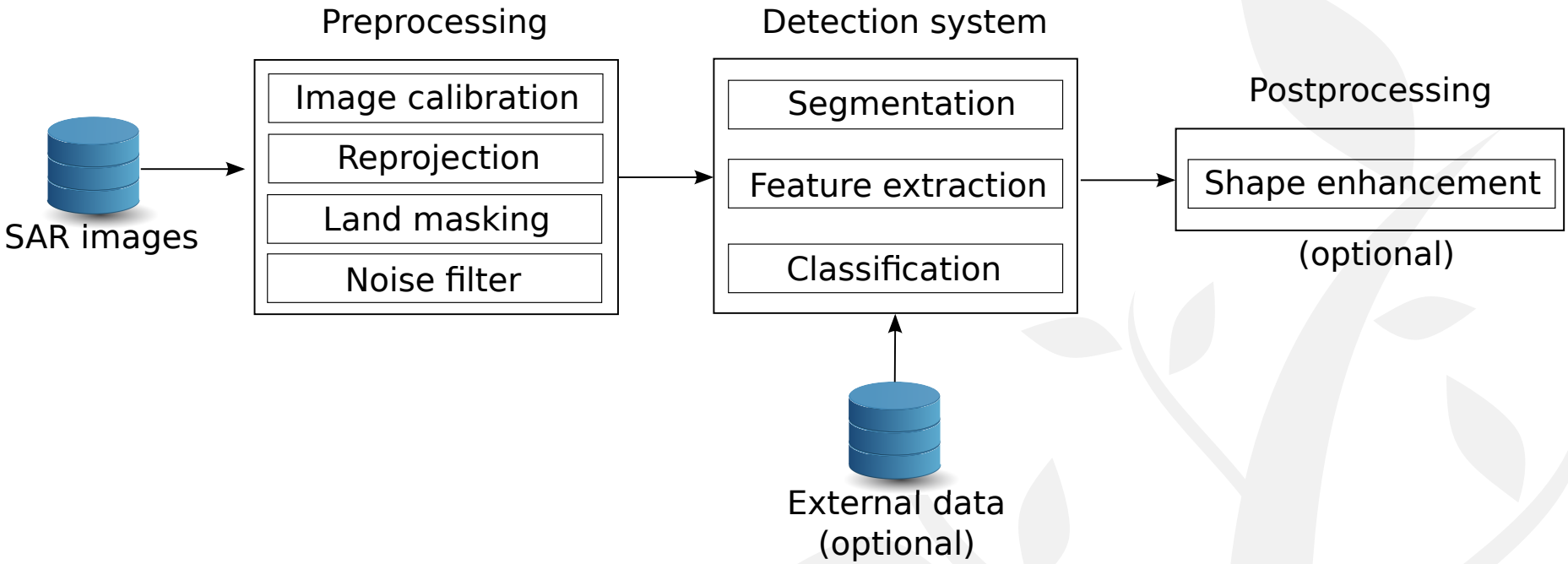


Prestige catastrophe
(Envisat, 17/11/2002)



Galician coast
(Envisat, 13/10/2008)

Sentinazos
Automatic detection system



- Segmentation: oil spill candidates are highlighted from the image background
- Feature extraction: segmented spots are analyzed to get their feature vectors
- Classification: each feature vector is labeled either as look-alike or oil spill

Sentinazos
Automatic detection system

- Dark spot segmentation
  - Typically, hundreds of candidates are segmented
- Feature selection/extraction
  - Features should have a good discriminatory power
  - Computationally intensive phase. Each candidate is independently examined
  - A concise and relevant set of features is crucial for obtaining a remarkable system accuracy

Sentinazos
Automatic detection system

- Feature selection/extraction
  - Literature shows many feature set alternatives
  - Typically, features are selected through the research experience
  - There are few systematic studies related to the feature selection and its influence in the classification phase

Index

Feature selection

Main objective

- Obtaining a concise and relevant subset of features to improve and to speed up the oil spill detection systems
- The validation of systematic ways to improve the oil spill feature selection

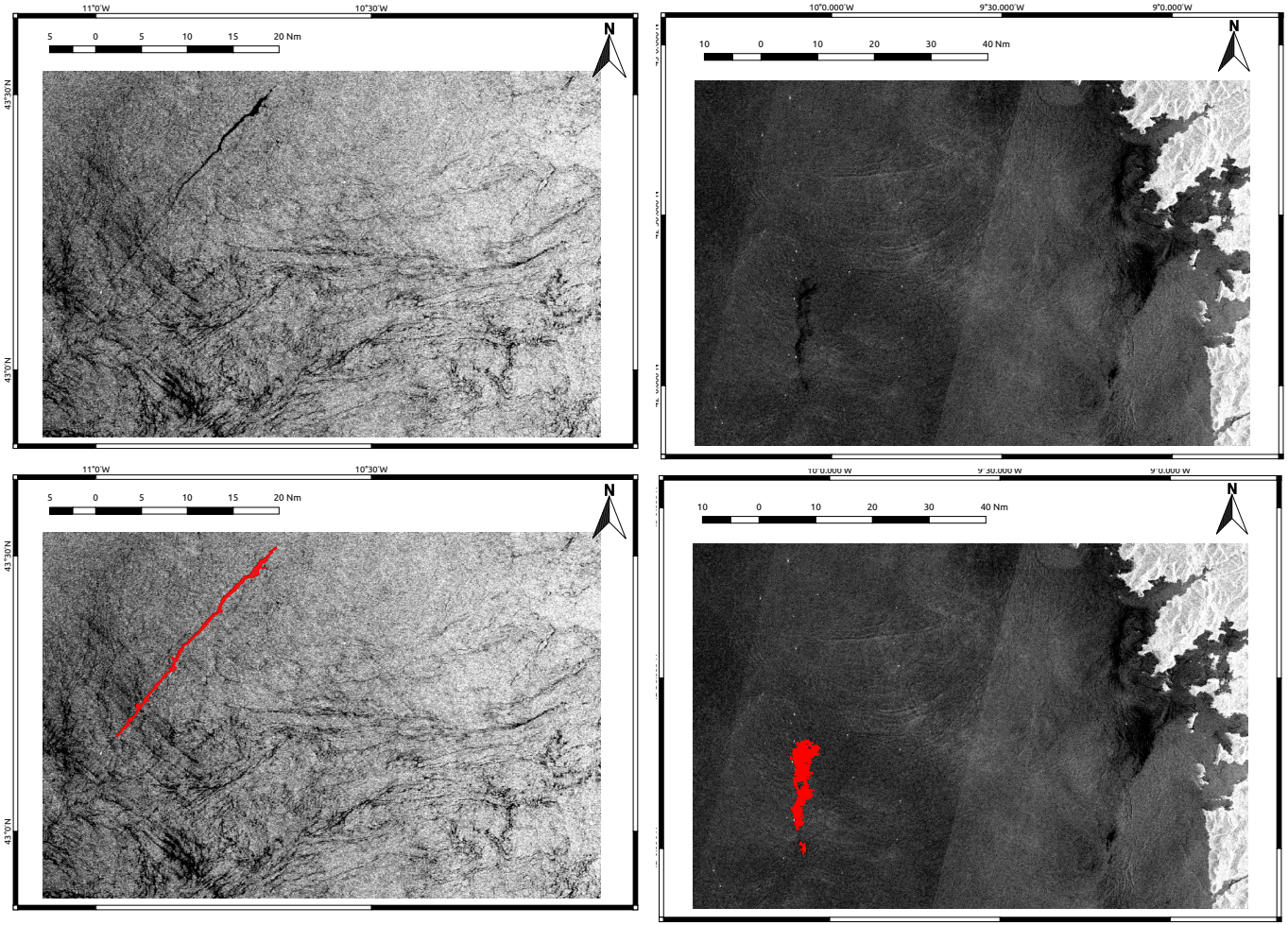
Feature selection

Methodology

- Dark spot segmentation
  - ROI: Galician coast (FTSS)
  - Oil spill database composed of elements obtained via both SASEMAR aircraft missions and EMSA reports
    - 47 spillages and 45 look-alikes
  - Database components were manually segmented for preserving their shape characteristics

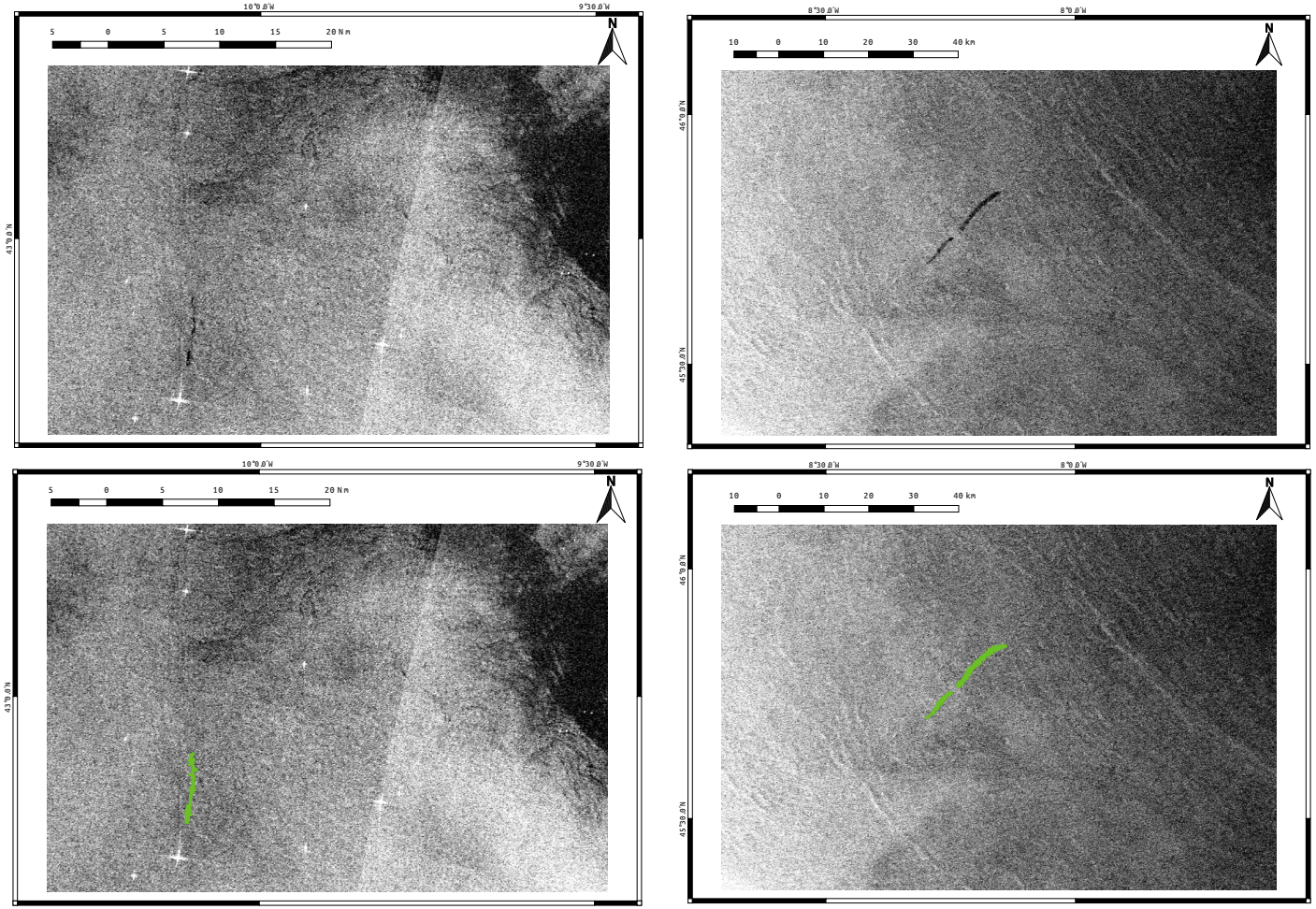
Feature selection Methodology

- Dark spot segmentation: oil spill examples



Feature selection Methodology

- Dark spot segmentation: look-alike examples



Feature selection

Methodology

- Feature extraction
  - Large feature vector based on a set of outstanding oil spill detection studies
    - 52 features - 141 components
    - Features from 3 main categories: geometrical, textural and physical

Feature selection
Methodology

- Feature selection methods
  - Preprocessing techniques within the machine learning field
  - Main goals
    - Discarding the irrelevant features
    - Improving the performance of the classifiers
  - Three main categories of feature selection methods
    - **Filters**: they make use of the general aspects of the data (i.e. entropy, correlation, etc.) to decide which subsets of features to keep
    - **Wrappers**: they are based on the performance of a specific prediction method to decide which subsets of features to select, usually retaining those which help improve the classifier performance
    - **Embedded methods**: they perform the feature selection stage as part of the learning process of a classifier

Feature selection

Methodology

- Selected feature selection methods
  - **Correlation-based feature selection (CFS)**: it selects subsets of features highly correlated with the class label but, at the same time, containing features uncorrelated with each other
  - **Consistency-based filter**: it measures the inconsistency of the feature subsets. A subset is inconsistent if there are at least 2 instances with the same feature values but with different class labels
  - **Information Gain**: it ranks all the features according to the information gain between a feature and a class
  - **ReliefF**: the success of this filter algorithm lies on the assumption that a good feature should be able to differentiate between samples from different classes whilst having the same value for samples from the same class
  - **SVM-RFE (Recursive Feature Elimination for Support Vector Machines)**: embedded method that trains a SVM classifier with the whole set of features and then it discards the least important ones

Feature selection

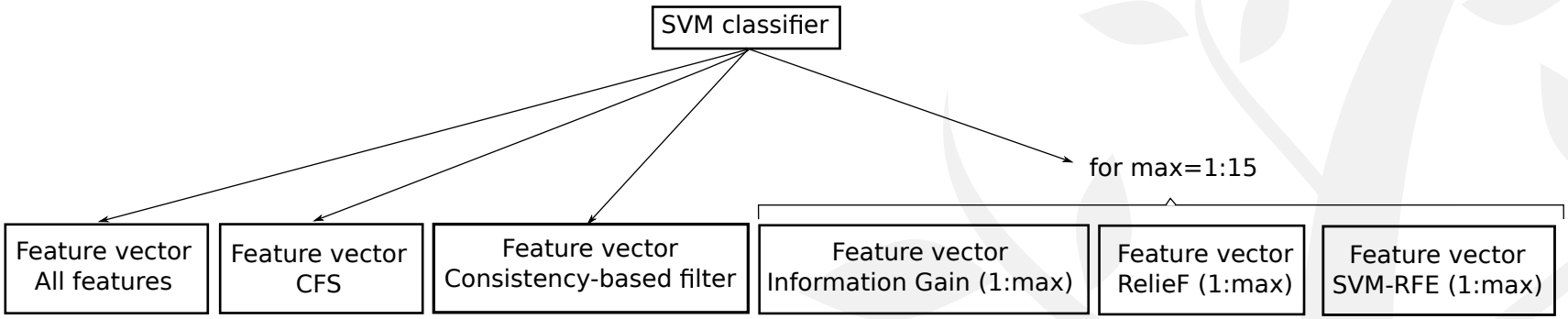
Methodology

- Implementation of the feature selection methods
  - Correlation-based feature selection (CFS): 8 features
  - Consistency-based filter: 6 features
  - Information Gain, ReliefF and SVM-RFE: 15 ranked features

Feature	CFS	Consistency-based filter	Information Gain	ReliefF	SVM-RFE
P/A				13	
S				6	
Sw				10	
Hu_1	X	X	9		
Hu_2			10		
Asm				7	5
L/W					14
N				9	8
Rs	X	X	1	2	1
Mr			2	1	3
Osd				4	
Crsd			12	11	
Opm			13		
Bpm			11		
Opm/Bpm			15	8	4
RISDI				3	9
RISDO			14	13	
Sc	X	X	3	5	2
Vas					13
H				12	
EFD_7	X	X	8		12
EFD_{11}					11
EFD_{23}	X	X	6		
EFD_{25}					15
EFD_{34}					6
EFD_{37}	X		4		7
EFD_{66}					10
EFD_{67}	X		5		
EFD_{79}	X	X	7		

Feature selection Methodology

- Training the classifier
  - Support vector machine (SVM) classifier
  - 48 feature vectors were checked



Feature selection Results

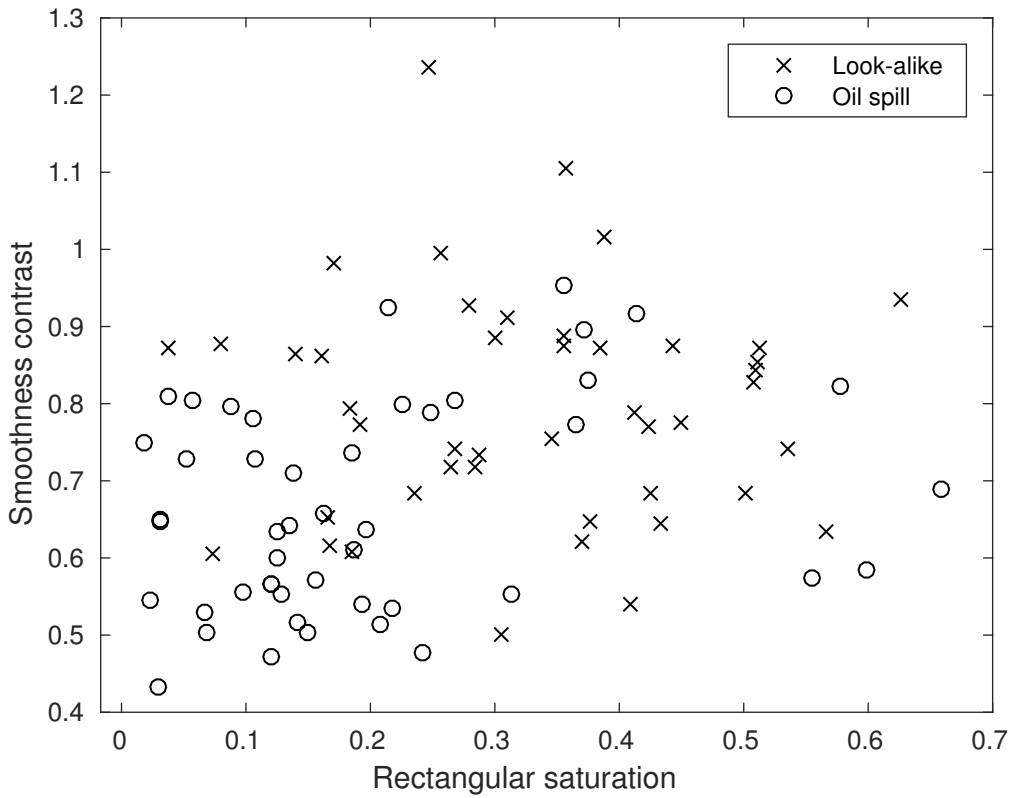
Feature selection method	Features	Accuracy (%)			Sens.	Spec.	Prec.	κ (%)
		OS	LA	Total				
None	141	80.00	62.50	70.97	0.8	0.63	0.67	42.24
CFS	8	60.0	75.00	67.74	0.6	0.75	0.69	35.15
Consistency based filter	6	73.33	87.50	80.65	0.73	0.88	0.85	61.09
Information Gain	5	66.67	81.25	74.19	0.67	0.81	0.77	48.12
ReliefF	9	66.67	87.50	77.42	0.67	0.88	0.83	54.51
SVM-RFE	**6**	**80.00**	**93.75**	**87.1**	**0.8**	**0.94**	**0.92**	**74.06**

- Sensitivity: the classifier ability to correctly identify those dark spots that are oil spills
- Specificity: the classifier ability to correctly identify those dark spots that are look-alikes
- Precision: the proportion of the true oil spills against all the positive results
- Cohen kappa statistic (κ): it measures the agreement between the true class and the classifier class excluding the probability of agreement by chance

Feature selection

Results

Feature name	CFS	Consistency-based filter	Information Gain	ReliefF	SVM-RFE
Rectangular saturation (Rs)	X	X	X	X	X
Smoothness contrast (Sc)	X	X	X	X	X
Ratio of the power to mean ratios (Opm/Bpm)			X	X	X
Marking ratio (Mr)			X	X	X
Asymmetry (Asm)				X	X
Elliptic Fourier Descriptor (EFD_{34})					X



Feature selection Results

- Comparative with previous oil spill detection systems

Reference	Features	OS	LA	Acc. (%)			κ (%)	Notes
				OS	LA	Total		
SVM-RFE	**6**	**47**	**45**	**80.00**	**93.75**	**87.1**	**74.06**	
[Topouzelis and Psyllos, 2012]	9	34	45	Unknown	Unknown	84.4	Unknown	
[Topouzelis et al., 2009]	10	34	45	85.29	84.44	84.81	69.2	Feature selection work
[Solberg et al., 1999]	11	71	6,980	94.37	98.92	98.88	62.41	
[Solberg et al., 2007]	13	37	12,110	78.37	99.36	99.3	40.29	
[Brekke and Solberg, 2008]	13	41	12,245	92.68	89.74	89.75	5.08	
[Fiscella et al., 2000]	14	80	43	85.00	67.44	78.86	53.01	Compound probability classification.
				92.5	51.16	78.05	47.49	Mahalanobis classification.
[Guo and Zhang, 2014]	50	222	863	79.27	98.07	94.12	81.39	Feature selection work. The study did not use an independent validation set

Feature selection Results

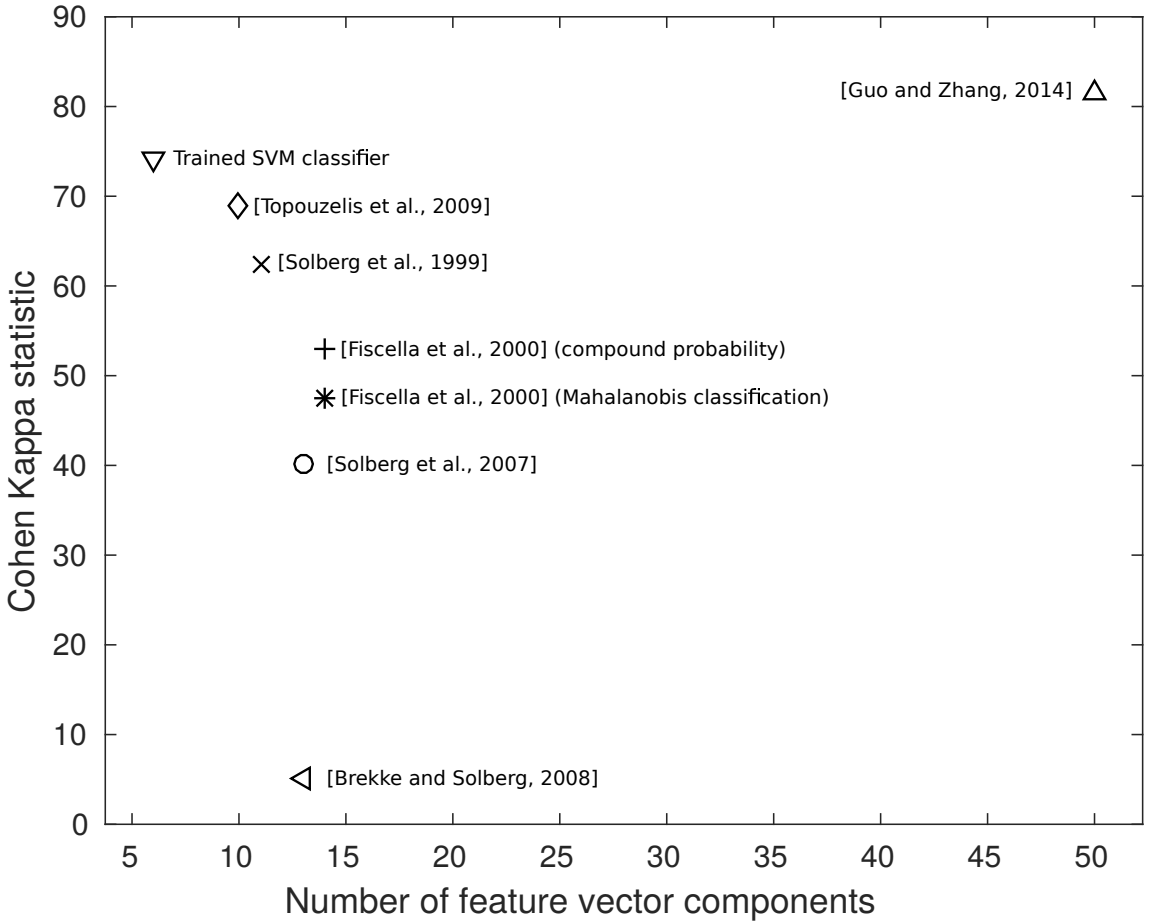


Figura : The κ statistic and the number of feature vector components for the trained SVM Classifier (SVM-RFE) and the other evaluated oil spill detection systems.

Feature selection Results

- Comparative between SVM classifiers trained with different feature vector compositions and the same dataset

Reference	Sel. method	Features	Acc. (%)			Sens.	Spec.	Prec.	κ (%)
			OS	LA	Total				
Our classifier	SVM-RFE	6	80.00	93.75	87.1	0.8	0.94	0.92	74.06
[Topouzelis et al., 2009]	Genetic algorithms	10	60.00	87.50	74.19	0.6	0.88	0.81	47.90
[Guo and Zhang, 2014]	Differential evolution feature selection	50	66.67	68.75	67.74	0.67	0.69	0.67	35.42

Index

Infrastructure

- Infrastructure
  - Intel Core i7@3.60GHz (4 cores)
  - 16GB RAM
- Candidate feature extraction cost
  - All features: 9.3 sec. (average)
  - 6-input feature vector: 2.0 sec. (average)
  - 78.49 % faster with feature selection
- Typically,a SAR image is composed of hundreds of candidates (mainly look-alikes)

Index

Conclusions

- Five feature selection methods were applied to a large feature vector
- The SVM-RFE selected an efficient combination of 6 features
  - Accuracy of 87.1 % and κ of 74.06 %
- We obtained a concise and relevant set of features
  - Reduction of the number of features (141$\rightarrow$6)
  - Improvement of the classifier performance
  - Reduction of the feature extraction process time (78.49 % faster)
  - Improvement of the problem knowledge
    - The proposed feature set is mainly composed of geometrical features (5/6)

References

Brekke, C. and Solberg, A. H. (2008).
Classifiers and confidence estimation for oil spill detection in envisat asar images.
Geoscience and Remote Sensing Letters, IEEE, 5(1):65–69.

Fiscella, B., Giancaspro, A., Nirchio, F., Pavese, P., and Trivero, P. (2000).
Oil spill detection using marine sar images.
International Journal of Remote Sensing, 21(18):3561–3566.

Guo, Y. and Zhang, H. Z. (2014).
Oil spill detection using synthetic aperture radar images and feature selection in shape space.
International Journal of Applied Earth Observation and Geoinformation, 30:146–157.

Solberg, A. H. S., Brekke, C., and Husoy, P. O. (2007).
Oil Spill Detection in Radarsat and Envisat SAR Images.
IEEE Transactions on Geoscience and Remote Sensing, 45(3):746–755.

Solberg, A. H. S., Storvik, G., Solberg, R., and Volden, E. (1999).
Automatic detection of oil spills in ers sar images.
Geoscience and Remote Sensing, IEEE Transactions on, 37(4):1916–1924.

Topouzelis, K. and Psyllos, A. (2012).
Oil spill feature selection and classification using decision tree forest on sar image data.
ISPRS Journal of Photogrammetry and Remote Sensing, 68:135–143.

Topouzelis, K., Stathakis, D., and Karathanassi, V. (2009).
Investigation of genetic algorithms contribution to feature selection for oil spill detection.
International Journal of Remote Sensing, 30(3):611–625.

Thank you for your attention!

david.mera@usc.es