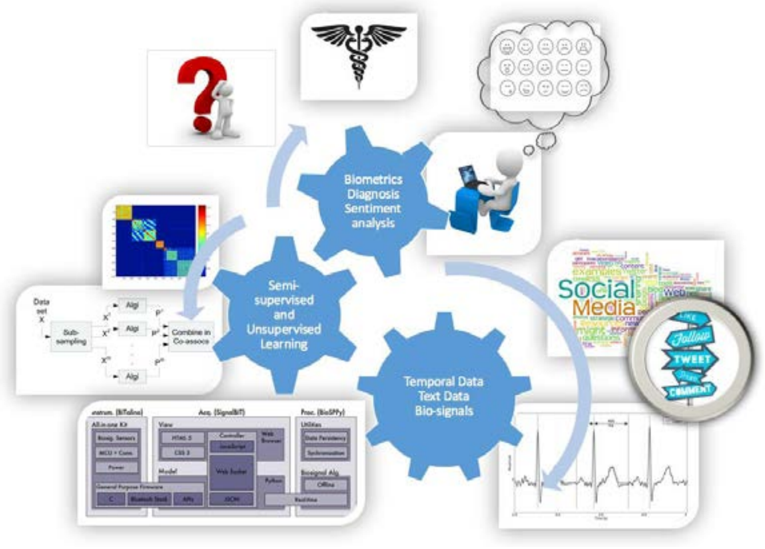
Learning from Sequences

This project proposes theoretical and algorithmic original contributions on exploratory data analysis techniques, namely on unsupervised and semi-supervised learning methods and ensemble methods, expanding these methods to the analysis and modelling of sequence data. Applications concern real world problems, namely automatic analysis of bio-signals, and text analysis from social networks.



Main Project Team	
Ana Luísa Nobre Fred	**PIA-Lx**
Mário Figueiredo	PIA-Lx
Hugo Silva	PIA-Lx
André Lourenço	PIA-Lx
Helena Aidos	PIA-Lx
André Martins	PIA-Lx/Priberam
Ana Priscila Alves	PIA-Lx
Carlos Carreiras	PIA-Lx
José Guerreiro	PIA-Lx
David Pereira Coutinho	PIA-Lx
Funding Agencies	
FCT/PTDC	**184,000€**
Start Date	01/04/2013
Ending Date	30/09/2015
Indicators	
Journal Papers	10
Conference Papers	38
Patents	1
Concluded MSc	5
Conference proceedings (as editor)	5

Two Main Publications

A. Lourenço, S. Bulo Bulo, N. Rebagliati, A. L. N. Fred, M. A. T. Figueiredo, M. Pelillo, **Probabilistic consensus clustering using evidence accumulation**, Machine Learning, Vol. 98, No. 1-2, pp. 331 - 357, April, 2015

H. Silva, A. Lourenço, A. L. N. Fred, N. Raposo, M. Aires-de-Sousa, **Check Your Biosignals Here: A New Dataset for Off-the-Person ECG Biometrics**, Computer Methods and Programs in Biomedicine, Vol. 113, No. 2, pp. 503 - 514, February, 2014

PROJECT WEBPAGE URL
https://www.it.pt/Projects/Index/1774

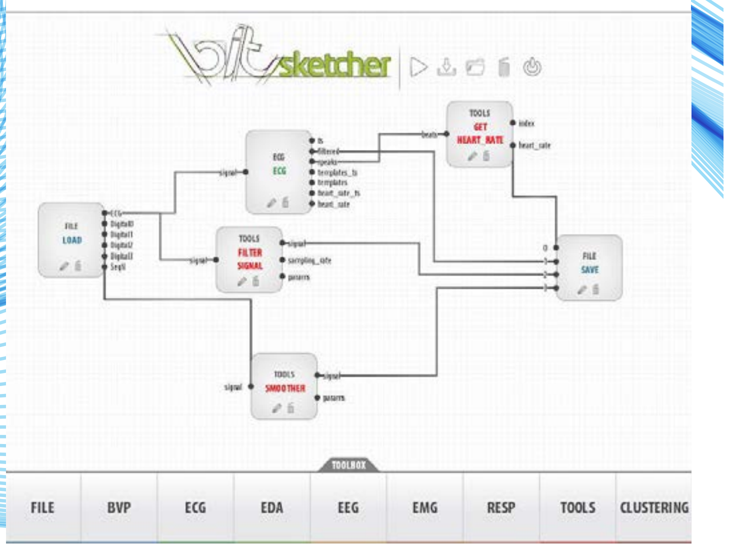


Fig. 1 Snapshot of the BITSketcher platform.

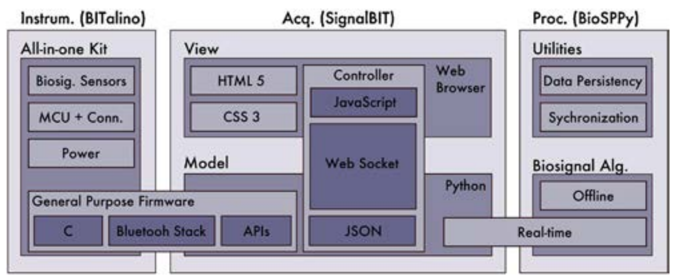


Fig. 2 Bio-signals Igniter Toolkit.

GENERAL MOTIVATION AND OBJECTIVES

The key idea of semi-supervised learning, and specifically constrained clustering, is to exploit both a priori information and unlabeled data to organize data into sensible groups. A priori information can be provided as labelled data or as constraints that one wishes to enforce or just encourage. Cluster ensemble methods, a research area in clustering introduced a decade ago, has gained increasing attention by the scientific community, both from theoretical and application domain perspectives. The Evidence Accumulation Clustering (EAC) paradigm, connecting the concepts of pairwise similarity and probability in unsupervised learning, has proven of great impact.

The main goals of this project are: 1) to advance the state-of-the-art on approaches for unsupervised and semi-supervised learning, with emphasis on cluster ensemble methods; 2) to expand these pattern recognition methods to the analysis and modeling of sequence data, corresponding either to sequences of symbols, such as in text, or sequential data in which observations are ordered in time, namely time series.

CHALLENGE

With the advent of the information society and the ubiquity of technologies and media for the production and retrieval of information, people increasingly confront a problem that seemed impossible only a few decades ago: too much information. Examples of this phenomena range from text and multimedia data (web pages, social media, scientific articles, ...) to biological signals. Automated analysis, structuring and summarization of relevant information thus becomes of utmost importance. We addressed this challenge by proposing innovative, state-of-the-art and scalable solutions in the areas of machine learning, pattern recognition and signal processing, focusing on semi-supervised and unsupervised learning methods, with main applications in bio-signal analysis (including biometrics, emotion assessment and diagnosis) and social media text analysis (including trend detection and sentiment analysis).

WORK DESCRIPTION AND ACHIEVEMENTS

The project puts forward novel contributions to the area of semi-supervised and unsupervised learning, extending research on ensemble and constrained clustering in terms of temporal data analysis and probabilistic assignment. The project addressed the combination of co-occurrences, Bayesian and information-theoretical approaches at several levels: feature extraction and dimensionality reduction; generation and combination of clusterings; and validation of clustering results. Scalability and higher order dissimilarity data representations were also addressed.

Concerning the processing and automatic analysis of bio-signals, we emphasise the development of a toolbox in Python with main signal processing and clustering methods – BioSpy (https://github.com/PIA-Group/BioSPPy) – and a web-based platform for easy development of signal processing and analysis systems by simple manipulation of functional block primitives – BITSketcher. Developed algorithms and proposed methodology were applied on physiological data for person identification (biometrics), emotion assessment, and clinical diagnosis.

The other main application area addressed trend detection and sentiment analysis from text from social media. In this context theoretical and algorithmic contributions were achieved, including syntactic and semantic parsing, a decoding algorithm with applications in natural language processing, new algorithms for transferring models across languages, in a weakly supervised learning paradigm, with application in domain adaptation for sentiment analysis and cross-lingual text classification and opinion mining.