

REFRAME Project



<http://www.reframe-d2k.org/>

Rethinking the **E**ssence, **F**lexibility and **R**eusability of **A**dvanced **M**odel **E**xploitation

José Hernández-Orallo
(Universitat Politècnica de València, Spain)

Madrid, 19 March 2015

Outline

- ▶ Project information
- ▶ Scientific background
- ▶ Challenges and potential impact
- ▶ Workplan, activities and results
- ▶ Management
- ▶ Dissemination and Value
- ▶ Conclusions

Project title and dates

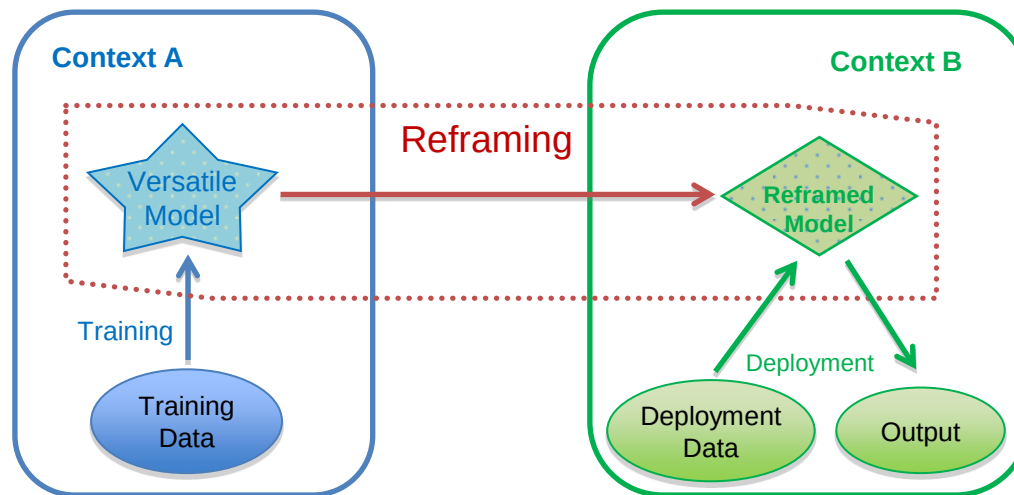
- ▶ REFRAME project
 - Rethinking the Essence, Flexibility and Reusability of Advanced Model Exploitation
- ▶ Dates:
 - START:
 - Officially: 1st October 2012 (kick-off meeting, Bristol)
 - Practically: 1st April 2013
 - END:
 - 31 March 2016 (6 months extension granted)

Consortium

- ▶ **BRIS: University of Bristol / Intelligent Systems Laboratory**
 - Peter Flach, Meelis Kull, Reem Al-Otaibi
- ▶ **STRAS: University of Strasbourg / LSIIT**
 - Nicolas Lachiche, Agnes Braud, Chowdhury Farhan Ahmed, Clement Charnay
- ▶ **VAL: Universitat Politècnica de València / DSIC**
 - Jose Hernandez-Orallo, Cesar Ferri, Maria-Jose Ramirez-Quintana, Adolfo Martínez-Usó

Scientific background (1)

- ▶ What is “reframing”?
 - “process of applying an existing [machine learning/data mining] model to the new operating context by the proper transformation of inputs, outputs and patterns”.



Scientific background (2)

► Examples:

- “Model predicting sales in Strasbourg for the following week may fail in Bristol for next Wednesday. The operating context has changed in terms of location as well as resolution of the inputs and outputs.”
 - Use a general model that can be applied at different locations and resolutions.
- “20% of the population will have an uncommon disease in the following ten years. The concept and frequency of *uncommon disease* changes.”
 - A general model is defined in terms of the background knowledge (e.g., “uncommon disease”).
 - If the background knowledge changes, the predictions of the model change.

Scientific background (3)

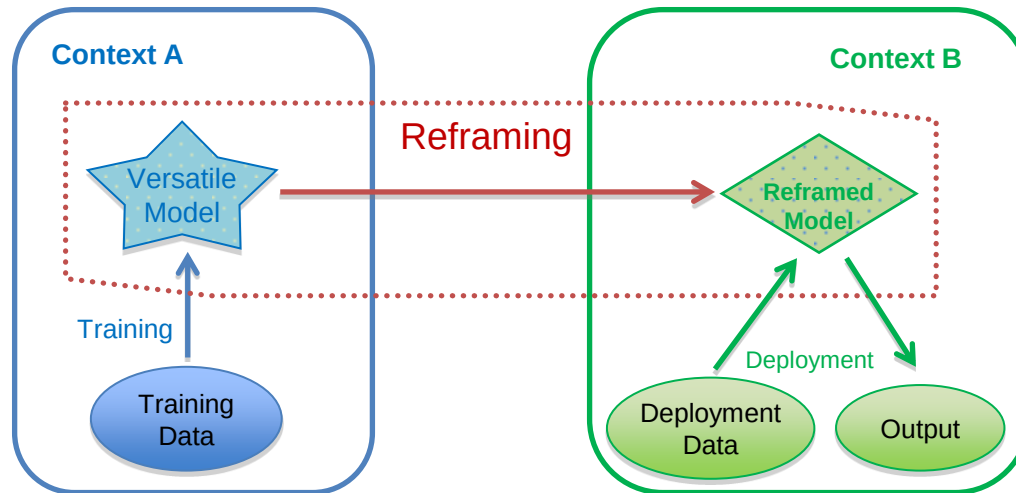
► What is a “context”?

◦ Examples:

1. Costs (output, inputs, labelling, etc.).
 2. Changes in data distribution.
 - Both 1 and 2 are integrated in tools such as ROC analysis.
 3. Changes and incompleteness in background knowledge.
 - A model can depend on the background knowledge.
 4. Granularity of feature (input) space or output space.
 - Multi-dimensional approach. Input features and output can be hierarchical.
 5. More or fewer (or different) attributes, missing data.
 6. Constraints
-
- One of the goals of the project is to generalise the notion of “operating context”.
 - So this list is becoming larger during the project.

Scientific background (4)

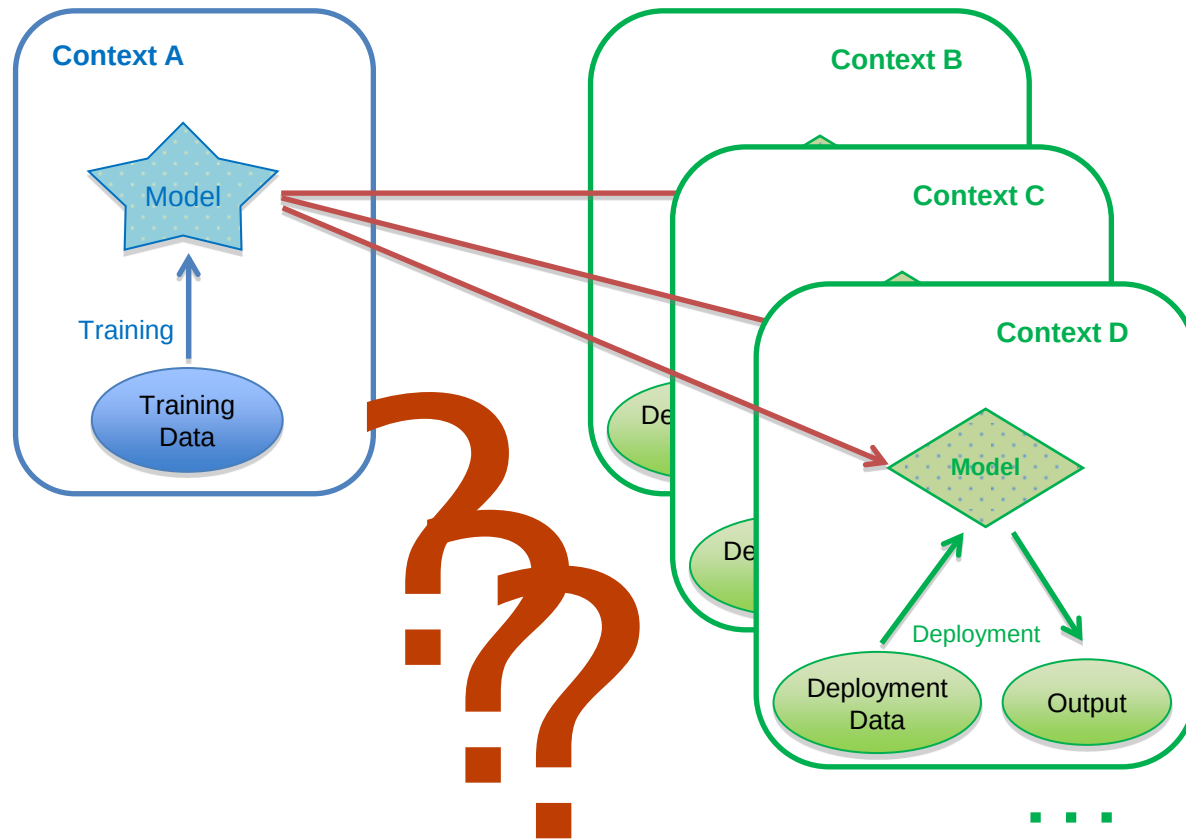
► “Versatile models”:



- The key issue is the construction of a “**versatile model**”, either directly or by “enrichment” of an existing model.
 - The versatile model may be more general (in structure, feature handling, output handling, knowledge dependencies, etc.) than really needed for context A.

Scientific background (5)

- Systematic (not a one-to-one transfer):



Challenges and potential impact (1)

- ▶ Usual D2K process:
 - Models are trained in a context.
 - This commonly results in overfitting to the training context.
 - Restricts model applicability whenever the training conditions change.
 - Models must be discarded and retrained repeatedly.

Inefficient and unreliable process.

Challenges and potential impact (2)

- ▶ We challenge that process:
 - We do not want to build a set of models for a range of operating contexts, or a very general, but inflexible model.
 - We aim at building **versatile models** that can be properly deployed in a range of operating contexts.

Any advance in this regard constitutes an important innovation and can have a strong impact on the way models are trained and deployed.

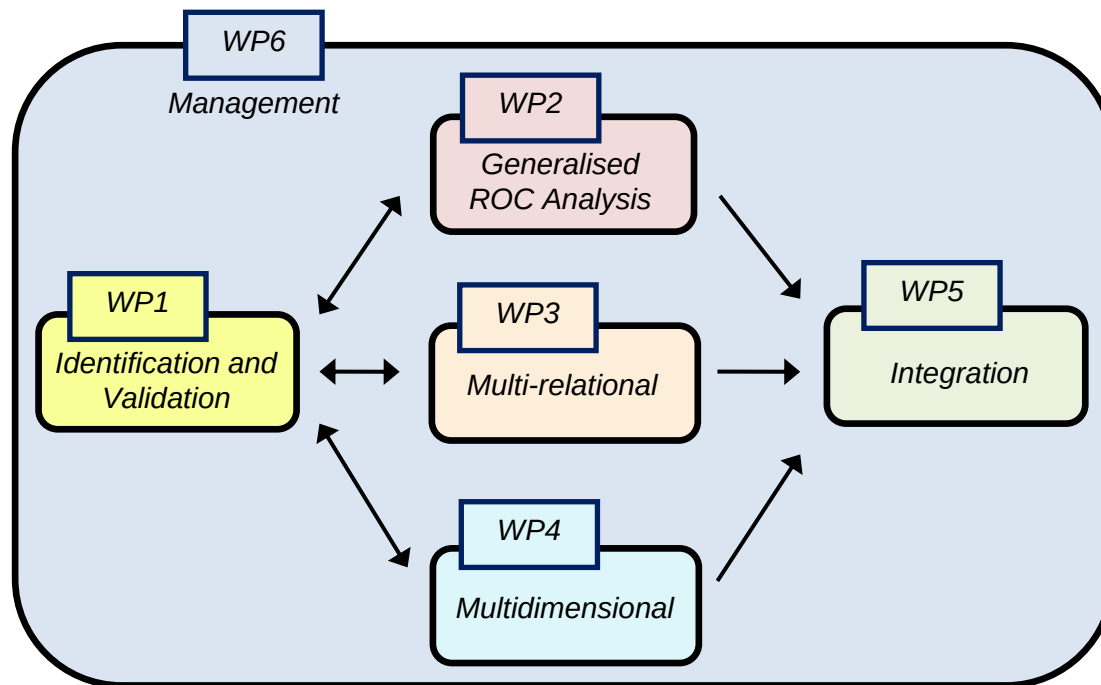
Challenges and potential impact (3)

- ▶ The purpose is to ensure “model reuse”.
 - Not the reuse of parts of old, existing models
 - The use of a general, versatile model for each possible context
 - Appropriate reframing procedures to apply it to any possible context.
 - Because of its long-life, the models can be robust and implemented in a more resource-efficient way.

Workplan, activities and results (1)

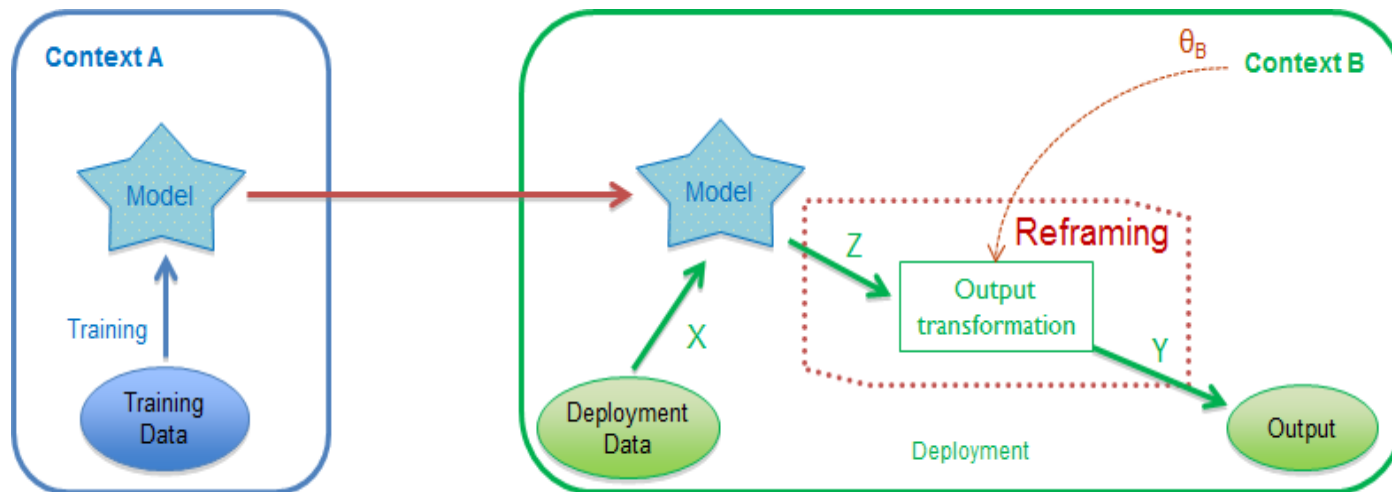
► Overview

- Six work packages:



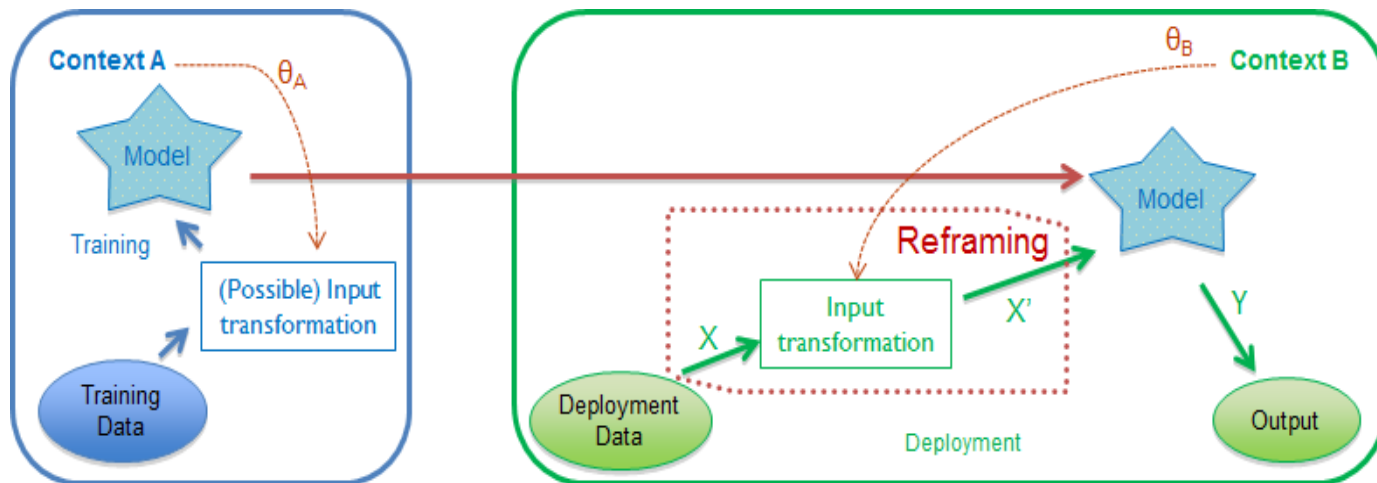
Workplan, activities and results (5)

- ▶ Characterisation of reframing
 - Output reframing:



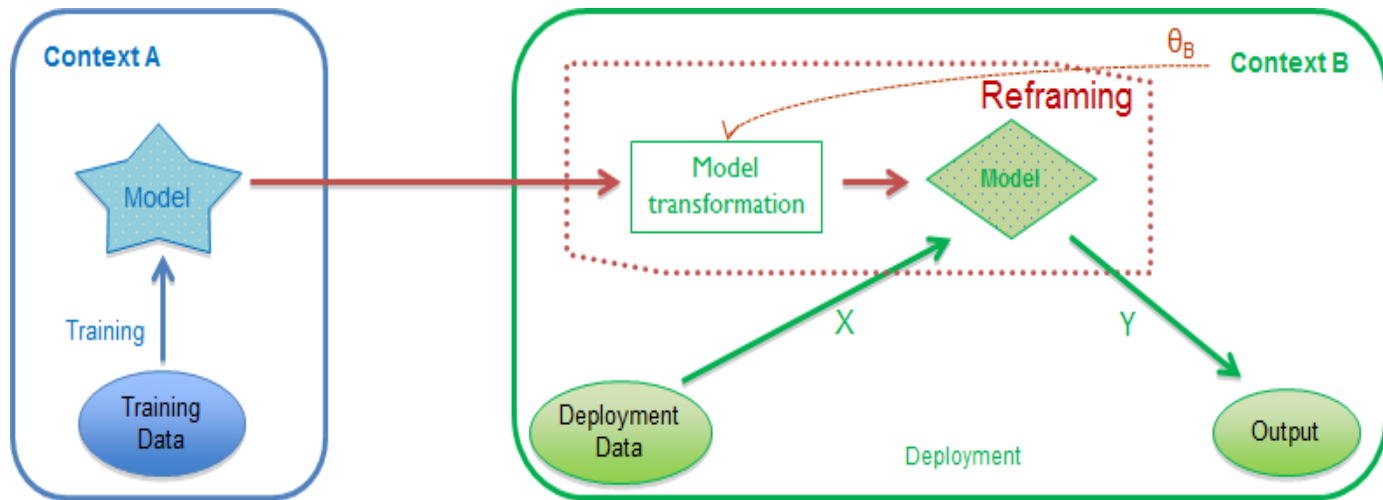
Workplan, activities and results (6)

- ▶ Characterisation of reframing
 - Input reframing:



Workplan, activities and results (7)

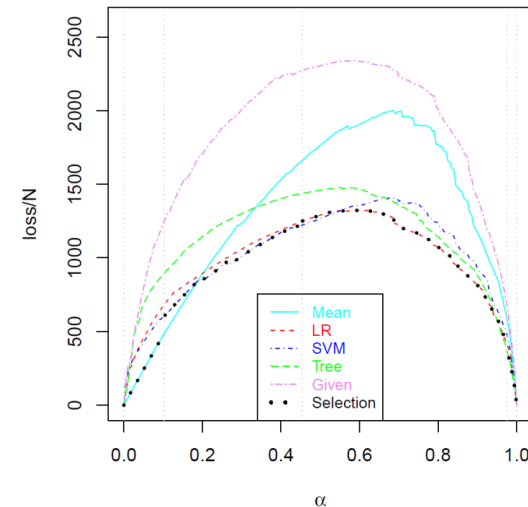
- ▶ Characterisation of reframing
 - Structural reframing:



Workplan, activities and results (8)

► New results for:

- Threshold methods and context plots:
 - In classification.
 - Cost contexts and context plots for regression.
 - Multilabel classification.
 - Multiclass: reliability estimation and ECOC.
- Input reframing:
 - Covariate shift: input attributes: univariate and multivariate.
 - Complex aggregates with relational features.
- Output reframing:
 - Aggregative quantification in the multidimensional case.
 - Aggregations and disaggregation in multidimensional contexts.
- Structural reframing
 - Versatile decision trees: using attribute statistics in the splits



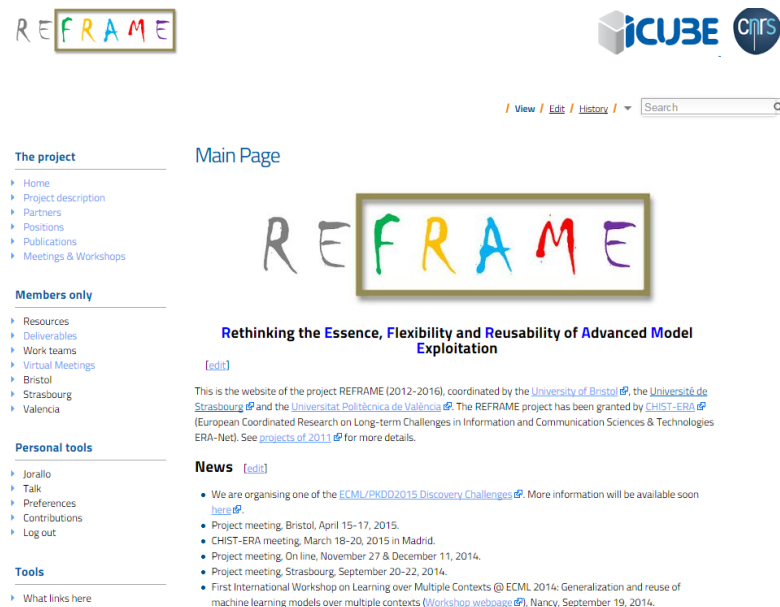
Workplan, activities and results (8)

► New results for:

- Contexts about information quality:
 - Dealing with uncertainty about the operating context.
 - Partial information about the context.
 - Definition and use of different noise contexts.
- Different types of data shift:
 - In relational data.
 - In graphs.
 - Identification of 12 patterns of shift (6 of them novel).
- Change of task:
 - Transformation of regression models into classification models.

Management (1)

- Website/Wiki:
<http://www.reframe-d2k.org/>



- Shared documents (links of the deliverables for evaluators).

Management (2)

- ▶ Internal project meetings:
 - Bristol, October 2012 (Kick-off)
 - Brussels, March 2013
 - Valencia, June 2013
 - Strasbourg, October 2013
 - Valencia, March 2014
 - Strasbourg, September 2014
 - Bristol, April 2015
- ▶ Along with several other one-day and two-day virtual meetings, mailing lists, etc.
- ▶ Research stays
- ▶ Analysis of new funding opportunities (H2020, ...)

Dissemination and Value (1)

- ▶ Publications (<http://reframe-d2k.org/Publications>)
 - Journals:
 - Journal of Machine Learning Research
 - Machine Learning
 - Pattern Recognition
 - Data Mining and Knowledge Discovery (1 + 1*)
 - ACM Transactions on Knowledge Discovery from Data
 - IEEE Transactions on Knowledge and Data Engineering*
 - Expert Systems with Applications*

(*) At different stages of the reviewing process.

Dissemination and Value (2)

- ▶ Publications (<http://reframe-d2k.org/Publications>)
 - Conferences:
 - International Conference on Inductive Logic Programming (2)
 - IEEE International Conference on Tools with Artificial Intelligence (2)
 - IEEE 13th international Conference on Machine Learning and Applications (2).
 - European Conference on Machine Learning / Practice of Knowledge Discovery in Databases (2)
- ▶ Also participation in workshops, summer schools, conference chairs, etc.

Dissemination and Value (3)

- Publications (<http://reframe-d2k.org/Publications>)
 - ECML PKDD 2014 Best paper award!
(<http://ecmlpkdd2014.loria.fr/conference/awards/>)



Conference

Slides
Awards
Photos

General Information

Registration
Important Dates
Organization
Program Committee
Invited Speakers
Industrial Track Invited Speakers
Guest Editorial Board
Sponsorship Opportunities
Sponsors
Discovery Challenges
Past Conferences
ECML PKDD 2015
ILP 2014
Frequently Asked Questions

Program

Schedule
Workshop list
Tutorial list

Awards

Best Paper Award ECMLPKDD 2014

● Reliability maps: A tool to enhance probability estimates and improve classification accuracy
(Meelis Kull, Peter Flach)

Authors

Meelis Kull, University of Bristol
Peter Flach, University of Bristol

Abstract

We propose a general method to assess the reliability of two-class probabilities in an instance-wise manner. This is relevant, for instance, for obtaining calibrated multi-class probabilities from two-class probability scores. The LS-ECOC method approaches this by performing least-squares fitting over a suitable error-correcting output code matrix, where the optimisation resolves potential conflicts in the input probabilities. While this gives all input probabilities equal weight, we would like to spend less effort fitting unreliable probability estimates. We introduce the concept of a reliability map to accompany the more conventional notion of calibration map; and LS-ECOC-R which modifies LS-ECOC to take reliability into account. We demonstrate on synthetic data that this gets us closer to the Bayes-optimal classifier, even if the base classifiers are linear and hence have high bias. Results on UCI data sets demonstrate that multi-class accuracy also improves.

Dissemination and Value (4)

- ▶ 1st Workshop on "Model Generalisation and Reuse over Multiple Contexts" at ECML/PKDD 2014 in Nancy.
 - <http://ecmlpkdd2014.loria.fr/program/workshop-list/>
 - <http://users.dsic.upv.es/~flip/LMCE2014/>

17/03/2015

LMCE 2014: First International Workshop on Learning over Multiple Contexts @ ECML 2014



LMCE 2014 # First International Workshop on Learning over Multiple Contexts @ ECML 2014

Generalization and reuse of machine learning models over multiple contexts

Nancy, France, 19 September 2014

A workshop held in conjunction with the [ECML PKDD 2014, Nancy, France, 15-19 September 2014](#)

Call for Papers

Adaptive reuse of learnt knowledge is of critical importance in the majority of knowledge-intensive application areas, particularly when the context in which the learnt model operates can be expected to vary from training to deployment. In machine learning this has been studied, for example, in relation to variations in class and cost skew in (binary) classification, leading to the development of tools such as ROC analysis to adjust decision thresholds to operating conditions concerning class and cost skew. More recently, considerable effort has been devoted to research on transfer learning, domain adaptation, and related approaches.

- ▶ 1st Workshop on "Model Generalisation and Reuse over Multiple Contexts" at ECML/PKDD 2014 in Nancy.
 - The program committee was selected to cover all those areas, with 18 members from 8 countries: Brazil, China, France, Japan, Portugal, Spain, UK and USA.
 - 2 invited talks, 13 regular papers and 3 work-in-progress papers.
 - Around 40 people participated in the day-long workshop.
- ▶ A second edition proposed for ECML/PKDD 2015 in Porto.

Dissemination and Value (5)

► Discovery Challenge at ECML/PKDD 2015, Porto.

- <http://www.ecmlpkdd2015.org/discovery-challenges>



[Home \(/\)](#) / [Global Info \(/global-info\)](#) / [Discovery Challenges](#)

DISCOVERY CHALLENGES

» Challenge #1 - Model reuse with bicycle rental data

Model reuse with bicycle rental data

Brief Description

Adaptive reuse of learnt knowledge is of critical importance in the majority of knowledge-intensive application areas, particularly when the context in which the learnt model operates can be expected to vary from training to deployment. This challenge therefore focuses on model reuse and context change.

The challenge is carried out in the framework of historical bicycle rental data obtained from several European cities. Bicycles are continuously taken from and returned to rental stations across the city. The data consists of time series describing hourly availability of bikes at each station; information on weather and (local) holidays is also provided. The challenge motivation is based on the fact that, while we may have had the opportunity to learn and tune good models for old stations with historical data, we do not always have the same amount of data for new stations. With that in mind, participants will receive, in addition to limited data for the new stations, a large number of trained models for old stations. The task will be to make predictions (several hours ahead) with regard to the number of bikes available for these new stations and within the next months. This situation fluctuates considerably depending on the time of year, the station's location, etc. The key point here is that by using models from other stations that have been learnt from data spanning more than one year, better predictions can be made for the new stations. In the end, this challenge aims at promoting the reusability of models rather than retraining a different model again and again each time the context changes.

◦ Model reuse with bicycle rental data.

- Possible collaboration with company J.C. Decaux.
- Models will be provided with PMML, the standard for data mining model exchange (data mining group:

<http://www.dmg.org/>)

Dissemination and Value (6)

- ▶ Software packages:
 - Open-source software integration.
 - Integration of different tools (data mining, OLAP, etc.)
 - New tools for model exchange (PMML).
 - Possible software exploitation by companies.

Conclusions

- ▶ Project as a short-term goal.
- ▶ More ambitious long-term goals:
 - Establish a solid and stable research effort to re-interpret the whole process of knowledge discovery from data.
 - Consolidate a research cluster in Europe as a catalyst for the area of machine learning evaluation and reuse worldwide.
- ▶ Dissemination and obtaining additional funding is key to build this cluster and make it sustainable.