DATA SCIENCE MEI/1

University of Beira Interior

Department of Informatics

Hugo Pedro Proença,

hugomcp@di.ubi.pt, **2021/22**

Meta Learning and Meta Mining

- Current data mining tools are characterized by an enormous number of algorithms that can be possible used for a specific task
- However, there is an evident lack of guidelines to select the right method according to the nature of the problem under analysis.
- Producing such guidelines is a primary goal by the field of **meta-learning**;
  - The goal is to understand the interaction between the mechanism of learning and the concrete contexts in which that mechanism is applicable.
- The field of meta-learning has seen continuous growth recently, with new developments in the construction of practical model-selection assistants, task-adaptive learners, and the proposal of solid conceptual frameworks.

Source: Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil and Carlos Soares. Using Meta-Learning to Support Data Mining. International Journal of Computer Science and Applications, Vol. I, No. 1, pag. 33-45, 2004.

Meta Learning and Meta Mining

- ❑ It should be noted that Meta-learning differs from base-learning in the scope of the level of adaptation;
  - ❑whereas learning at the base-level is focused on accumulating experience on a specific learning task (e.g., credit rating, medical diagnosis, mine-rock discrimination, fraud detection, etc.)
  - ❑Learning at the meta-level is concerned with accumulating experience on the performance of multiple applications of a learning system.
- ❑Meta-learning solves important problems in data mining (DM) tools, particularly in the problems of “Classification” and “Regression”.
- ❑Without some assistance, mode selection can be an obstacle to end-users (Data Scientists)
  - ❑The availability of **too many** models to proceed on a trial-and-error basis (e.g., by measuring accuracy via some re-sampling technique such as n-fold cross-validation), can be an intractable problem.
- ❑A solution to this problem is attainable through the construction of meta-learning systems. These systems provide automatic user guidance by mapping a particular task to a suitable model (or combination of models).

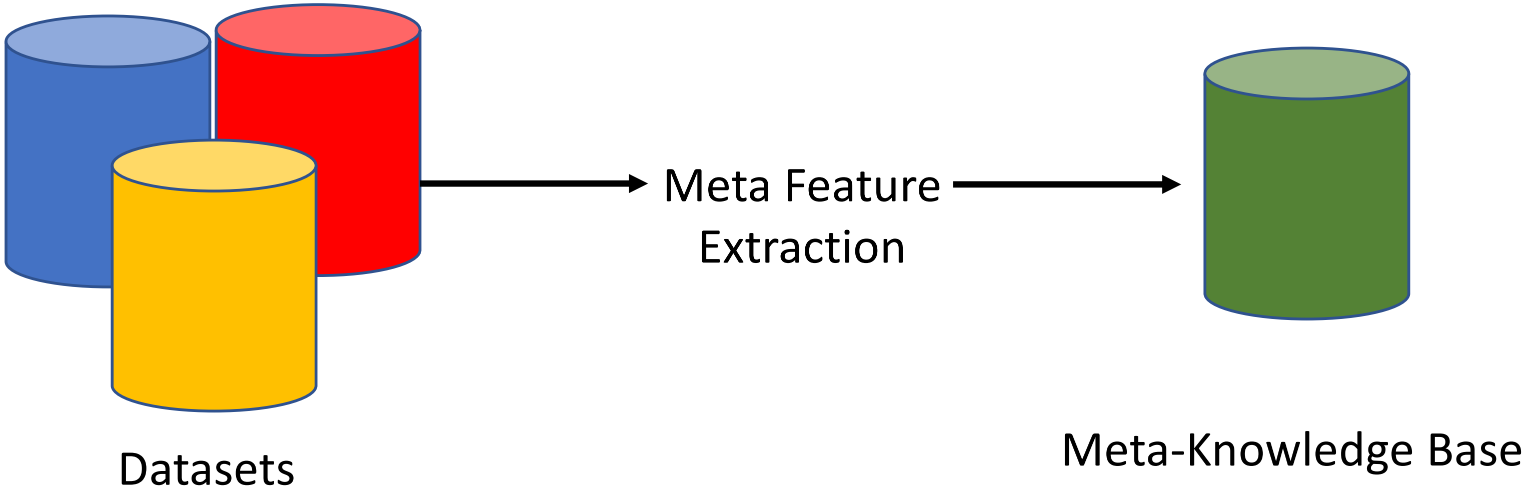
Meta Learning and Meta Mining

- ❑ A second contribution of Meta Learning tools is to profit from the repetitive use of a predictive model over similar tasks. The successful application of models in real-world scenarios requires continuous adaptation to new needs.
- ❑Rather than starting afresh on new tasks, one would expect the learning mechanism itself to re-learn, taking into account previous experience
- ❑Meta-learning systems can help control the process of exploiting cumulative expertise by searching for patterns across tasks.
- ❑Conceptually, a meta-learning system can be divided into two modes of operation
  - ❑**Knowledge Acquisition**;
  - ❑**Knowledge Advisory**.

Source: Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil and Carlos Soares. Using Meta-Learning to Support Data Mining. International Journal of Computer Science and Applications, Vol. I, No. 1, pag. 33-45, 2004.

Meta Learning: Knowledge Acquisition

- ❑ During this step, the main goal is to learn about the learning process itself.
- ❑ We assume that the input to the system consists of datasets of examples.
  - ❑ For each dataset, the meta-learning system extracts dataset characteristics or meta- features.
  - ❑ The goal of this step is to gather information that transcends a particular domain of application.

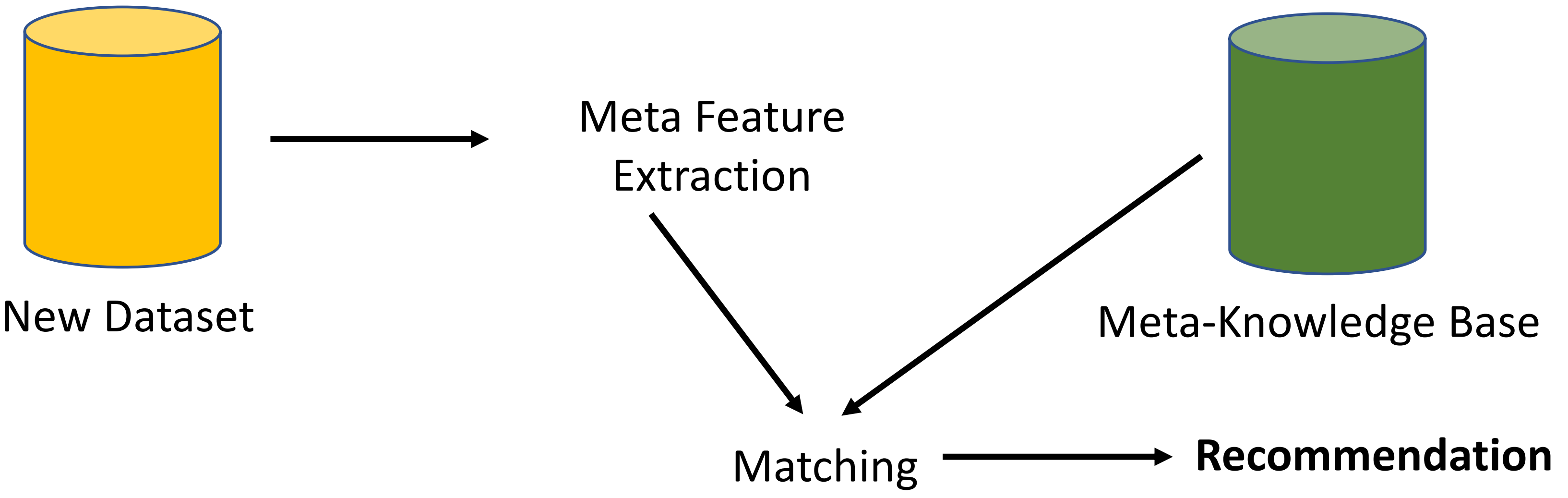


Meta Learning: Knowledge Acquisition

- ❑The “Meta-Feature Extractor" looks for information that can be used to generalize to other example distributions.
- ❑During the Knowledge Acquisition phase, the learning techniques are not expected to exploit knowledge of previous results.
- ❑Statistics derived from different learning strategies (e.g., a classifier or combination of classifiers) and their performance may be used as a form of characterizing the task under analysis.
- ❑Information derived from the meta-feature generator and the performance evaluation module can be combined into a meta-knowledge base.
- ❑This knowledge base is the main result of the knowledge acquisition phase; it reflects experience accumulated across different tasks.

Meta Learning: Advisory Mode

- ❑During the advisory mode, the meta-knowledge previously acquired is used to configure the learning system in a manner that exploits the characteristics of the new data.
- ❑Meta-features extracted from the dataset are “matched” with the meta-knowledge base to produce a recommendation regarding the best available learning strategy.
- ❑At this point we move away from experimentation with the base learners to the ability to do informed model selection or combination of base learners.



Meta Learning: Advisory Mode

- ❑The resulting “Recommendation” can be composed of:
  - ❑**Best preprocessing techniques;**
  - ❑**Best parameters settings;**
  - ❑**Best model (model selection);**
  - ❑**Best combination of models;**
- ❑The effectiveness of the meta-learner increases as it accumulates meta- knowledge.
- ❑The lack of experience at the beginning of the learner's existence compels the meta-learner to use one or more learning strategies without a clear preference for any one of them; experimenting with many different strategies is time consuming.
- ❑However, as more training sets are examined, we expect the expertise of the meta-learner to dominate the process of deciding which learning strategy suits best the characteristics of the current problem.

Characterization of Datasets

- ❑An important component of any meta-learning system is to extract relevant information about the task under analysis.
- ❑The key idea is that high-quality dataset characteristics or meta-features provide some information to discriminate between the performance of a set of given learning strategies.
- ❑**Statistical and Information-Theoretic Characterization**
  - ❑Several works in dataset characterization have concentrated on extracting statistical and information-theoretic parameters estimated from the training set.
  - ❑Such measures include number of classes, number of features, ratio of examples to features, degree of correlation between the features and the target(s), the average class entropy and class-conditional entropy, skewness, kurtosis, signal–to-noise ratio.

Characterization of Datasets

- ❑**Model-Based Characterization**
  - ❑In addition to statistical measures, a different form of dataset characterization exploits the properties of the induced hypothesis as a form of representing the dataset itself. As an example, one can build a decision tree from a dataset and collect properties of the tree (e.g., nodes per feature, maximum tree depth, shape, tree imbalance, etc.). This can be seen as a way to characterize the dataset.
- ❑**Landmarking**
  - ❑The idea is to exploit information obtained from the performance of a set of simple learners (i.e., learning systems with low capacity) that exhibit significant differences in their learning mechanism. The accuracy (or error rate) of these landmarkers is used to characterize a dataset and identify areas where each of the simple learners can be regarded as an expert.
  - ❑Another idea related to landmarking is to exploit information obtained on simplified versions of the data (e.g. samples). Accuracy results on these samples serve to characterise individual datasets and are referred to as sub-sampling landmarks. This information is subsequently used to select an appropriate learning algorithm

Mapping Datasets to Models

- ❑In practice, the most important and practical use of meta-learning is the development of a mechanism that maps an input space composed of datasets to an output model space, composed of predictive models. Criteria such as accuracy, storage space, and running time can be used for performance assessment.
- ❑**Hand-Crafting Meta Rules**
    - ❑Using human expertise and empirical evidence, a number of meta-rules matching domain characteristics with learning techniques may be crafted manually.
    - ❑For example, in learning a decision tree, a heuristic rule can be used to switch from univariate tests to linear tests if there is a need to construct non-orthogonal partitions over the input space.

Mapping Datasets to Models

- ❑ **Learning at the Meta-Level**
  - ❑ After learning from several tasks, one can construct a meta-dataset where each element pair is made up of the characterization of a dataset (meta-feature vector) and a class label corresponding to the model with best performance on that dataset. A learning algorithm can then be applied to this well-defined learning task to induce a hypothesis mapping datasets to predictive models.
  - ❑ A variation to the approach above is to look at the neighborhood of a query in the space of meta-features. When a new query dataset is presented, the k-nearest neighbor instances (i.e., datasets) around this dataset are identified to select the model with best average performance.
  - ❑ Instead of mapping a task or dataset to a predictive model, a different approach consists of selecting a model for each individual query example. The idea is similar to the nearest neighbor approach: select the model displaying best performance around the neighborhood of the query example.

Mapping Datasets to Models

❑**Ranking Models**

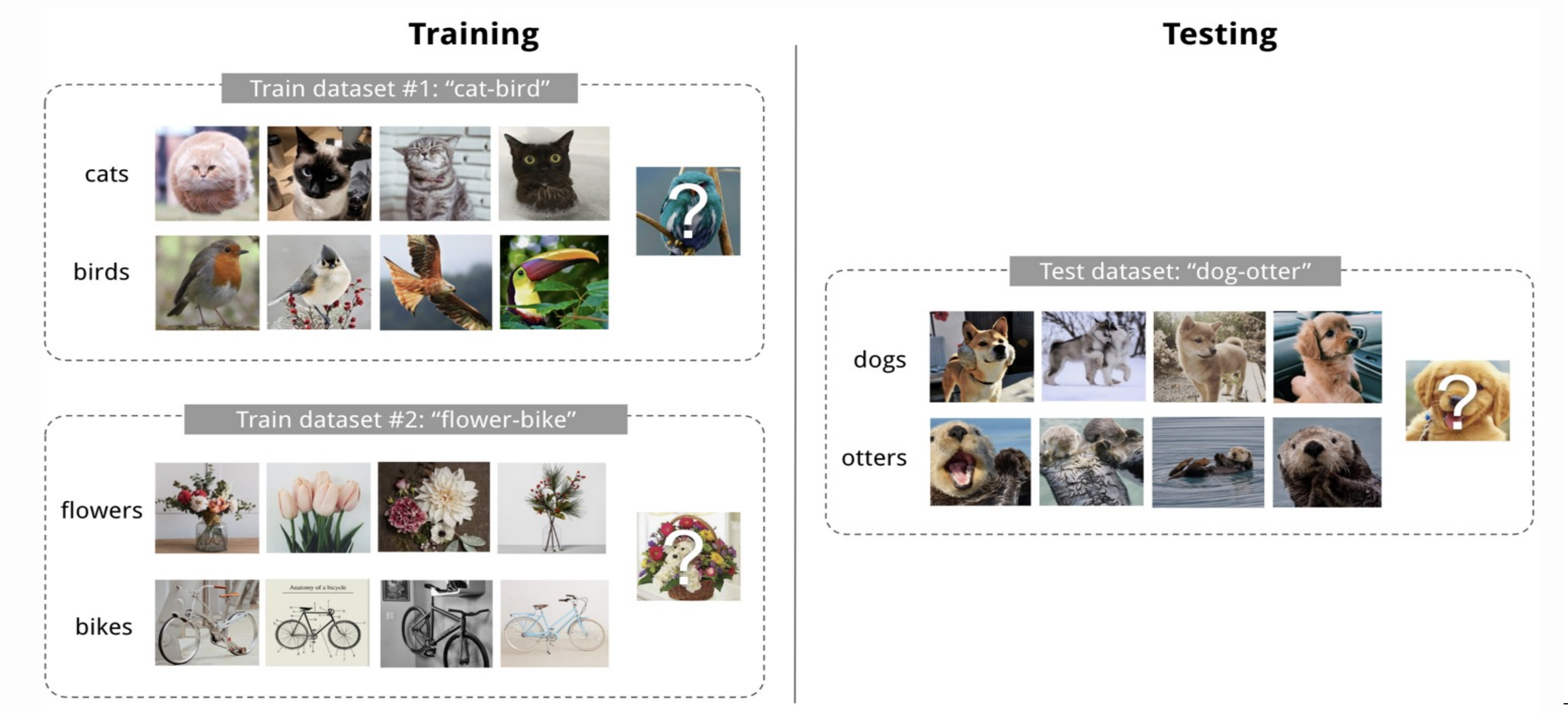
- ❑Instead of mapping a dataset to a single predictive model, one may also produce a ranking over a set of different models.
- ❑It can be argued that such rankings are more flexible and useful to users. In a practical scenario, the advice should not be limited to a single item; this could lead to problems if the suggested final model is unsatisfactory.
- ❑Rankings provide alternative solutions to users who may wish to incorporate their own expertise or any other criterion (e.g., financial constraints) into their decision-making process.

Combining Base-Learners

- ❑Another (and perhaps more popular) approach to meta-learning consists of learning from base-learners. The idea is to make explicit use of information collected from the performance of a set of learning algorithms at the base level; such information is then incorporated into the meta-learning process.
- ❑**Stacked Generalization**
  - ❑This process works under a layered architecture. Each of a set of base-classifiers is trained on a dataset; the original feature representation is then extended to include the predictions of these classifiers. Successive layers receive as input the predictions of the immediately preceding layer and the output is passed on to the next layer. A single classifier at the topmost level produces the final prediction.
  - ❑This is considered a form of meta-learning because the transformation of the training set conveys information about the predictions of the base-learners.

Meta Learning Example

- ❑Consider the following Learning and Testing Datasets.
  - ❑The initial Datasets contain information to discriminate between “Cats” and “Birds” (#1) and “Flowers and Bikes (#2)
  - ❑However, we are interested in discriminate between “Dogs” and “Otters”, having no access to enough corresponding learning data.


Training
Testing
Train dataset #1: “cat-bird”
cats
birds
?
Train dataset #2: “flower-bike”
flowers
bikes
?
Test dataset: “dog-otter”
dogs
otters
?

Meta Learning Example

- ❑One of the most popular solutions is the concept of "Siamese CNN", which is composed of two twin networks and their outputs are jointly trained on top with a function to learn the relationship between pairs of input data samples.
- ❑The twin networks are identical, sharing the same weights and network parameters. In other words, both refer to the same embedding network that learns an efficient embedding to reveal relationship between pairs of data points.
- ❑The Siamese network is trained for a **verification task** for telling whether two input images are in the same class. It outputs the probability of two images belonging to the same class.
- ❑This kind of networks is not learning the concepts of "Cat" or "Bird", but instead it is learning the concept of "Equal"/"Different".
- ❑The assumption is that the learned embedding can be generalized to be useful for measuring the distance between images of unknown categories.