Meta-learning

Włodzisław Duch,
Department of Informatics, Nicolaus Copernicus University, Poland,
School of Computer Engineering, Nanyang Technological University, Singapore
wduch@is.umk.pl (or search “Duch W”)

Synonyms

Definition

Meta-learning methods are aimed at automatic discovery of interesting models of data. They belong to a branch of Machine Learning that tries to replace human experts involved in the Data Mining process of creating various computational models learning from data. Given new data and description of the goals meta-learning systems should support decision making in classification (Classification), regression (Regression, Statistics), association tasks, and/or provide comprehensible models of data (Rule-Based Methods). The need for meta-learning came with availability of large data mining packages such as Weka that contain hundreds of components (data transformations) that may be connected in millions of ways, making the problem of optimal model selection exceedingly difficult. Meta-learning algorithms that “learn how to learn” and guide model selection have been advanced in Statistics, Machine Learning, Computational Intelligence and Artificial Intelligence fields.

Characteristics

Learning from data, or understanding data requires many pre-processing steps, selection of relevant information, transformations and classification methods. Without *a priori* knowledge about a given problem finding an optimal sequence of transformations is a great challenge. According to the “No-free lunch theorem” (Duda et al. 2001) no single method may outperform others in all situations. The best recommendation is to try many different approaches and transfer knowledge gained by solving other problems. Meta-learning techniques help to select or create optimal predictive models and reuse previous experience from analysis of other problems, relieving humans from most of the work and realizing the goal of computer programs that improve with experience (Brazdil et al. 2009; Jankowski et al. 2011). These methods are designed to automatize decisions required for application of computational learning techniques.

Related machine learning subjects include transfer learning (concerned with learning a number of related tasks together), lifelong learning, multi-task, cross-domain and cross-category learning, and transfer learning from unlabeled data, called self-taught learning. Meta-learning and meta-reasoning is of great interest for the artificial intelligence community (Vilalta et al. 2004; Anderson & Oats 2007). It is also of great relevance to computational biology, where hierarchical multi-task learning allows to transfer knowledge from one task to another, gaining experience from training predictive models on a large number of related organisms.

Meta-feature approach

Perhaps the simplest goal for meta-learning is to recommend the best method for a given data that consists of objects described by some features. This requires characterization of the dataset by meta-features, ability to estimate similarity of new data to already analyzed datasets, and a database of results obtained by various methods on these datasets. In the clusterization or in the supervised learning problems the role of such meta-learning systems is to recommend the use of methods that gave the best result on a similar dataset. This approach has been used in several projects (Brazdil et al. 2009) using statistical and information-theoretic measures gathered initially in the StatLog project (Michie et al. 1994) and the Metal project (Giraud-Carrier et al. 2004), such as the number of samples, features, classes, types of features, class entropy, average feature entropy, average mutual information, noise-signal ratio, outlier measure of continuous features, number of features with outliers. All these meta-features are quite easy to compute but without additional measures of data complexity quality of recommendations is very limited. Simple geometrical data complexity measures (Basu & Kam Ho, 2006) include measures of linear separability, Fisher's discriminant ratio, ratio of average intra/inter-class distance, volume of overlap region, fraction of instances on class boundary, and error rates for fast classifiers (called *landmarkers*) run on the whole or on the random subsets of data. Using such meta-descriptors classifiers or regression models are trained to rank the usefulness of available predictive methods for a given data selecting best approaches with some success.

The space of possible solutions generated by recommendation systems is restricted to already known types of algorithms and usually only a relatively small number of base algorithms is included in the ranking. Moreover, these methods do not recommend which pre-processing transformations or specific options of the base algorithms should be used.

Sometimes ensemble methods (committees of data models) are also acknowledged as meta-learning systems. Because they do not automatize decisions related to applications of learning models they should rather belong to the additional or post-learning stage of data analysis. Combination of data models leads to a very rough granularity of knowledge, thus exploring only a small subspace of all possible models. It does not facilitate insights into the hidden structures in the data, nor does it allow to explore alternative models. The key to greater flexibility is to use heterogeneous learning systems (Duch 2007), based on different primitive elements (similarity evaluation, neural transfer functions, tests used by decision trees etc.), combined with a complexity-guided search for optimal models.

Meta-learning as search for optimal models

The great challenge in data mining is to create flexible learning systems that can extract relevant information from data, transfer knowledge from past experience, and reconfigure themselves to find various interesting solutions for a given task (Duch 2007). Instead of a single learning algorithm designed to solve specialized problem, priorities should be set to define what makes an interesting solution, and a search for configurations of computational modules that automatically create algorithms on demand should be performed. This search

in the space of all possible models should be constrained by user priorities and should be guided by experience with solving problems of similar nature. With no prior knowledge about a given problem finding an optimal sequence of transformations may not be possible, it has to be learned by analyzing patterns of successful workflows for many types of models on many types of data.

Initial attempts to create meta-learning systems that search in the space of all possible models have been restricted to the similarity-based framework (Duch 2000; Duch et al. 2000) using only simple greedy search techniques. In case of classification problems probability p(C|X;M) of assigning class C to a vector X, given a classification model M, depends on adaptive parameters and procedures used in construction of the model. Starting from the simplest nearest-neighbor model extensions based on optimization of the number of neighbors, feature selection, feature weights, type of distance function, weighting of distances, selection of reference vectors, and other parameters and procedures that may improve the model, are systematically investigated. If the added complexity justifies gains in the quality of extended model it is accepted as the current one, otherwise the search stops. The workflows in this approach included only fixed order of search through admissible extensions, from the lowest to the highest computational complexity. The final methods could be presented as a network of transformations, and depending on the data it could implement all variants of the nearest neighbor methods, various artificial neural networks (Radial Basis Functions, Multilayer Perceptrons, Separable Functions Networks, Learning Vector Quantization), neurofuzzy models and many others (Duch and Grudziński 2002).

Meta-learning based on search for optimal composition of transformations defined on a set of objects requires several components:

- transformations specific for signals, images, texts, sequences or general measurements that act as filters extracting relevant information from raw data, creating useful support features that capture various aspects of data, including similarities of objects to known reference objects;
- general guiding principles that can be used to create higher-level object description, such as preservation of similarity, discovery of common factors and discriminative characteristics, high variance of features, independence of signals, local separability of large clusters, measures of covariance and correlation/dependency of targets on new descriptors, information transfer between different types of predictive models;
- schemes or templates defining promising workflows for compositions of transformations that create high-order description of objects, images of input data in enhanced spaces after all transformations;
- analyzing patterns of data distribution in these enhanced spaces;
- models of decision processes that can answer queries based on these patterns;
- learning associations between the type of data, workflows that create new image of data and decision processes;
- organization of search based on these associations.

This approach goes in many ways well beyond recommendations which method to use for a given data. First, it tries to compose new predictive methods from a series of transformations, and thus may create quite novel combination of transformations. They are discovered using general principles, for example looking for good prototypes of objects with specific characteristics or class labels, and transformations that preserve similarity to these objects. Such transformations create new features that have direct interpretation, features that are implicitly used by advanced machine learning methods such as kernel Support Vector Machines. Other types of features may include linear projections that create independent components, or projections that generate large clusters of objects of the same type (Maszczyk et al. 2010). Various algorithms may contribute useful features, for example decision trees may discover interesting combinations of several input features that can be used as higher-order features, and the nearest neighbor methods may provide features based on the similarity functions to good prototypes $z(X)=D(X,R)$.

Hierarchical methods for generation of information starting from raw features and add new support features that are discovered by different types of data models created on similar tasks, successively building more complex features that form the basis of enhanced feature spaces. General principles inspiring this approach are derived from complementarity of information processed by parallel interacting streams within hierarchical organization of information processing in the brain. In the enhanced feature space the goal of learning is to create image of the input data that can be directly handled by relatively simple decision processes. Usually such data image shows some patterns of data clusters. For example, problems with complex inherent logic may show clusters on a line with vectors that belong to alternating classes.

Transfer of knowledge is possible thanks to the fine granularity of knowledge representation at the level of transformations that generate new features. For meta-learning it may not be important that a specific combination of features proved to be successful in some task, but it is important that a specific transformation of a subset of features was useful for data of this type, or that distribution of patterns in the feature space had some characteristics that may be described by some specific data model and thus is easy to adapt to a new dataset of similar type. Such information allows for generalization of knowledge at the level of search patterns for a new composition of transformations, facilitating transfer of knowledge between different tasks. Resulting algorithms facilitate deep learning that may only be achieved by composition of several transformations, help to discover many alternative models, and also enable understanding of structures hidden in the data by logical, fuzzy and prototype-based rule discovery based on new features, and by visualization of the results of data transformations.

General search for models in enhanced feature spaces requires a very sophisticated search strategies. Some ideas for building such systems, controlling the process of composing transformations based on monitoring of model complexity, and learning useful workflows have recently been proposed (Grabczewski and Jankowski 2008), and the Intemi (Intelligent Miner) software based on these principles is under development. Availability of such software will change the future of data mining, enabling domain experts with little experience

in computational methods a deep analysis of their data that should lead to important discoveries.

Cross-references

Classification, Machine Learning
Regression, Statistics
Decision Rule, Machine Learning
Machine Learning
Artificial Intelligence
Confidence, Machine Learning
Support, Machine Learning
Knowledge Discovery, Machine Learning
Induction, Logics
Training Set, Machine Learning
Generalization Ability, Machine Learning
Bias-Variance Trade-Off
Entropy, Information Theory
Kolmogorov-Smirnov Distance
Information Gain
Overfitting, Machine Learning
Stability, Machine Learning
Weka, Machine Learning Tool
R, Data Analysis Tool

References

Anderson M.L, Oates T. (2007) A review of recent research in metareasoning and metalearning. AI Magazine 28: 7-16.

Basu M, Kam Ho T, eds. (2006) Data Complexity in Pattern Recognition. Springer.

Brazdil P, Giraud-Carrier C, Soares C, Vilalta R (2009) Metalearning: Applications to Data Mining. Cognitive Technologies. Springer.

Duch W (2000) Similarity based methods: a general framework for classification, approximation and association, Control and Cybernetics 29 (4): 937-968.

Duch W (2007) Towards comprehensive foundations of computational intelligence. In: Challenges for Computational Intelligence, Springer, pp. 261-316

Duch W, Adamczak R, Diercksen G.H.F. (2000) Classification, Association and Pattern Completion using Neural Similarity Based Methods, Applied Mathematics and Computer Science 10: 101-120.

Duch W, Adamczak R, Grąbczewski K (2001) A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Transactions on Neural Networks 12: 277-306

Duch W, Grudziński K (2002) Meta-learning via search combined with parameter optimization. In: Advances in Soft Computing, eds. L. Rutkowski and J. Kacprzyk, Springer, pp. 13-22.

Duch W, Setiono R, Zurada J.M (2004) Computational intelligence methods for understanding of data. Proc. of the IEEE 92(5): 771- 805

Duda R.O, Hart P.E, Stork D (2001) Patter Classification. J.Wiley & Sons, New York

Giraud-Carrier Ch, Vilalta R, Brazdil P. (2004) Introduction to the Special Issue on Meta-Learning, Machine Learning 54: 197-194

Grąbczewski K (2008) Meta-learning with Machine Generators and Complexity Controlled Exploration. Lecture Notes in Artificial Intelligence 5097: 545-555

Grąbczewski K, Duch W (2002) Heterogenous forests of decision trees. Lecture Notes in Computer Science 2415: 504-509

Jankowski N, Grąbczewski K, Duch W (2011) Meta-learning in Computational Intelligence. Springer (in print).

Maszczyk T, Grochowski M, Duch W (2010) Discovering Data Structures using Meta-learning, Visualization and Constructive Neural Networks, Studies in Computational Intelligence 262 :467-484, Springer.

Michie D, Spiegelhalter D.J, Taylor C.C (1994) Machine learning, neural and statistical classification. Elis Horwood, London

Vilalta R, Giraud-Carrier C.G, Brazdil P, Soares C (2004) Using meta-learning to support data mining. International Journal of Computer Science & Applications 1(1): 31–45

Vilalta R, Drissi Y (2002) A Perspective View and Survey of Meta-Learning. Artificial Intelligence Review 18: 77-95.

In: Encyclopedia of Systems Biology,
W. Dubitzky, O. Wolkenhauer, K-H Cho, H. Yokota (Eds.),
Springer 2011