

Multi-target prediction: A unifying view on problems and methods

Willem Waegeman · Krzysztof Dembczyński ·
Eyke Hüllermeier

Received: date / Accepted: date

Abstract Multi-target prediction (MTP) is concerned with the simultaneous prediction of multiple target variables of diverse type. Due to its enormous application potential, it has developed into an active and rapidly expanding research field that combines several subfields of machine learning, including multivariate regression, multi-label classification, multi-task learning, dyadic prediction, zero-shot learning, network inference, and matrix completion. In this paper, we present a unifying view on MTP problems and methods. First, we formally discuss commonalities and differences between existing MTP problems. To this end, we introduce a general framework that covers the above subfields as special cases. As a second contribution, we provide a structured overview of MTP methods. This is accomplished by identifying a number of key properties, which distinguish such methods and determine their suitability for different types of problems. Finally, we also discuss a few challenges for future research.

Keywords multivariate regression, multi-label classification, multi-task learning, pairwise learning, dyadic prediction, zero-shot learning, collaborative filtering

1 Introduction

In contrast to conventional supervised learning, where a single target variable needs to be predicted on the basis of a set of features describing an instance, multi-target prediction (MTP) is concerned with the simultaneous prediction of multiple target variables of possibly different type, such as binary, nominal, ordinal, or real-valued. Applications of multi-target

W. Waegeman
Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653,
B-9000 Ghent, Belgium
E-mail: willem.waegeman@ugent.be

K. Dembczyński
Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland
E-mail: kdembczynski@cs.put.poznan.pl

E. Hüllermeier
Department of Computer Science, Paderborn University, Pohlweg 49–51 33098 Paderborn, Germany
E-mail: eyke@upb.de

prediction are omnipresent in the digitalized society of the 21st century. Classical applications that are often studied in machine learning papers include image tagging in computer vision, document categorization in text mining, and product recommendation in online advertising. Besides, MTP problems arise in many other application domains as well. In medicine, one is interested in predicting several clinical outcomes of patients at the same time. In climate sciences, one would like to forecast extreme weather events for many regions in the world, which are related through complex physical and geological processes. In biology, one would like to unravel the different biological functions that a gene might express. In chemistry, one would like to know which molecules might be potential drugs to cure a given disease. In social networks, one intends to predict which users interact with a given user. In ecology, one constructs species distribution models, which describe the prevalence of different types of species in a given habitat.

Applications of that kind have resulted in novel research questions, and a need for developing specialized multi-target prediction methods. Often, several target variables are related to each other, for example because they obey certain constraints (for instance, if variables are positions in a ranking, they must be mutually exclusive), they correspond to nodes of a graph, or they provide evidence of parent-child relationships. In other situations, specific properties of the targets are known, such as molecular structures or feature representations. Obviously, knowledge of that kind could be used to improve predictive performance. However, in many applications, neither target relations nor representations are known a priori. Instead, they need to be discovered from the data. The main credo of research in MTP is the conviction that, compared to the most obvious approach of learning an individual prediction model for each target variable independently of the others, the exploitation of target dependencies will lead to better performance.

In this review-style article, we present a unifying view of MTP problems and methods. When speaking about multi-target prediction, machine learning researchers often allude to either multi-label classification, multi-output regression (a.k.a. multivariate regression in statistics), or multi-task learning problems. Moreover, if additional side information in the form of target relations or target representations is accessible, those three settings further extend to multi-target prediction scenarios that are known as dyadic prediction, hierarchical multi-label classification, and zero-shot learning in the literature. Besides, there is a close connection to matrix completion and network inference, which becomes obvious when representing the relationship between instances and targets in the form of a matrix or a graph. Despite strong commonalities, there is little interaction between the different sub-communities. Moreover, there are several problems that have been studied in different communities under different names. Sometimes there is even terminological confusion within the same community.

As a first contribution, we provide a structured overview of the multitude of MTP problems. To this end, we present a formal framework for multi-target prediction in Section 2. By identifying a set of characteristic properties, we subdivide MTP problems into a number of well-known settings. Particular attention will be paid to the formal definitions of problems that characterize the fields of multi-label classification, multivariate regression, multi-task learning, zero-shot learning, and matrix completion. We also discuss a number of related settings, such as structured output prediction and multi-class classification, and argue why those settings should not be covered by the umbrella of MTP.

Subsequently, in Section 3, we present a unifying view of MTP methods. Here, we intend to unravel a number of general mechanisms that are essential for obtaining state-of-the-art predictive performance in MTP. As will be seen, the applicability or usefulness of a method strongly depends on properties of the problem setting, such as whether or not side

information is available for targets, and how this side information looks like. Another important question is whether or not one intends to generalize to novel targets and/or instances.

The overall goal of this paper is to provide insight into the vast literature on MTP problems and methods, especially for readers who are new to the field. The paper is not a typical review paper, however, and does not lay claim on being comprehensive in this regard—in light of the breadth of the field, that appears to be an impossible endeavor. Instead, we intend to focus on some general principles that might be helpful in identifying the right approach for a given problem. In Section 4, we conclude with a couple of remarks and challenges for future research, and briefly discuss some important aspects of multi-target prediction that are less emphasized in this paper.

2 A unifying view on MTP problems

2.1 A general framework for multi-target prediction

In this section, we establish links between different MTP problems. We start by describing a general framework that covers both simple and more advanced MTP problems. Formally, our framework is defined as follows.

Definition 1 (Multi-target prediction) A multi-target prediction setting is characterized by instances $\mathbf{x} \in \mathcal{X}$ and targets $\mathbf{t} \in \mathcal{T}$ with the following properties:

- P1. A training dataset \mathcal{D} consists of triplets $(\mathbf{x}_i, \mathbf{t}_j, y_{ij})$, where $y_{ij} \in \mathcal{Y}$ denotes a score that characterizes the relationship between the instance \mathbf{x}_i and the target \mathbf{t}_j .
- P2. In total, n different instances and m different targets are observed during training, with n and m finite numbers. Thus, the scores y_{ij} of the training data can be arranged in an $n \times m$ matrix Y , which is in general incomplete, i.e., Y has missing values.
- P3. The score set \mathcal{Y} is one-dimensional. It consists of nominal, ordinal or real values.
- P4. The goal consists of predicting scores for any instance-target couple $(\mathbf{x}, \mathbf{t}) \in \mathcal{X} \times \mathcal{T}$.

The above definition accommodates the availability of side information for targets. To keep notation simple, we stick to vector representations in our definitions, i.e., we identify targets with feature vectors \mathbf{t} . We remark, however, that other types of side knowledge, such as structured representations or relations, could also be considered. The examples to be discussed below will make this point more clear.

In what follows, we show that various multi-target prediction problems are recovered as specific instantiations of the above framework. As already said, if side information for targets is available, we assume it can be encoded in the vector representation \mathbf{t} . If no side information is available, \mathbf{t} will be an uninformative vector (e.g., merely consisting of an identifying number).

2.2 Conventional multi-target prediction settings

We start by explaining what we call conventional multi-target prediction problems. For those problems, side information for targets is normally not available. Multi-label classification, multivariate regression, and multi-task learning are the most well-known subfields of machine learning that can be mentioned as examples. Multivariate regression and multi-label classification consider the simultaneous prediction of multiple binary or real-valued targets,

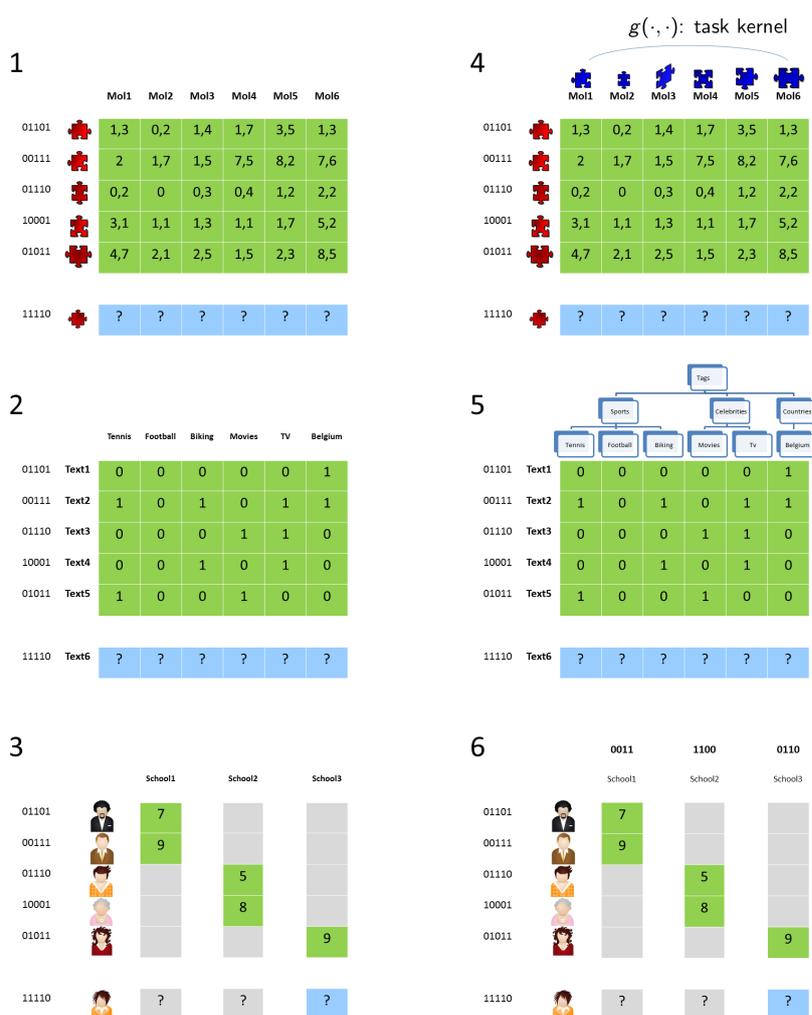


Fig. 1 A visualization of six prototypical multi-target prediction problems: (1) drug design as an example of a multivariate regression problem, (2) document categorization as an example of a multi-label classification problem, (3) student grading as an example of a multi-task learning problem, (4) drug design with a similarity measure for targets, (5) document categorization with a hierarchy for targets, (6) student grading with target features. See Examples 1 to 6 in the text for more details.

respectively (Dembczyński et al, 2012; Tsoumakas and Katakis, 2007). Multi-task learning then unifies those subfields, and further extends them to problems where not all targets are relevant for all instances (Caruana, 1997). To make this point clearer, we discuss three prototypical examples.

Example 1 Consider the problem of predicting whether a protein will bind to a set of experimentally developed small molecules. This is an important application in the area of drug design, where machine learning methods can help in screening interesting novel chemical compounds. Using multi-target prediction methods, high-potential compounds can be

selected and analyzed further during (more rigorous) wet-lab experimentation. Figure 1-1 illustrates the data that might be generated in an application of that kind. For a set of training proteins (first five rows, in green), the binding affinity with a set of small molecules (represented as columns) has been verified experimentally. Furthermore, we assume that we have additional information about the proteins available, in the form of a feature representation or a more structured representation such as a graph (shown as red puzzle pieces in the picture). One can use this data to train a multi-target prediction model that is able to predict, on the same set of small molecules, binding affinities for new proteins that have not yet been analyzed in the wet-lab. As experimental verification of binding affinities delivers continuous scores that represent binding strength, one arrives at a multivariate regression problem.

Example 2 In the area of document categorization, assigning appropriate category tags to documents is important but laborious and time consuming. Hence, one often applies multi-target prediction methods to automate this tagging process. As shown in Figure 1-2, one might have asked a human annotator to provide a training dataset with relevant tags being assigned to a set of documents. Using a feature representation of those documents, such as a bag-of-words representation, one can then train a multi-target prediction method that will be able to assign tags to new documents outside the training dataset. More than one tag might be relevant for a particular document, so automated document categorization boils down to solving a multi-label classification problem.

Example 3 We consider the application of predicting student marks in the final exam for a typical high-school course, let's say mathematics. Forecasting those marks at an early stage, prior to the start of a course, might be useful to give students advice with respect to study directions and perspectives. Using historical student records that might be stored by schools, one can train a multi-target prediction method that estimates the exam marks of newly entering students. To this end, one would typically construct a dataset that looks like Figure 1-3, where columns represent different courses or different schools. If one aims for estimating the marks for different courses or different schools simultaneously, one ends up with solving a multi-task learning problem.

Examples 1–3 discuss basic multi-target prediction settings in which no additional knowledge about the targets is known. Example 1 is a multivariate regression problem, as one aims to predict the values of continuous variables, whereas Example 2 is a multi-label classification problem with binary target variables. Example 3 depicts an application that can only be tackled with multi-task learning methods. As students usually attend only one school, only one label will be observed for per student. The other labels are unknown and typically irrelevant—they are represented by grey cells in Figure 1-3. Likewise, in the prediction phase, it is meaningful to restrict predictions to the school of the student, or at least a subset of schools for which the student is considering an enrolment. In most multi-task learning problems, only one task or a subset of tasks is relevant for a given instance. We therefore introduce the following formal definitions to include multivariate regression, multi-label classification, and multi-task learning in our framework.

Definition 2 (Multivariate regression) A multivariate regression problem is a specific instantiation of the general framework, which exhibits the following additional properties:

- P5. The cardinality of \mathcal{T} is m . This implies that all targets are observed during training.
- P6. No side information is available for targets. Without loss of generality, we can hence assign the numbers 1 to m as identifiers to targets, such that the target space is $\mathcal{T} = \{1, \dots, m\}$.

P7. The score matrix Y has no missing values.

P8. The score set is $\mathcal{Y} = \mathbb{R}$.

Definition 3 (Multi-label classification) A multi-label classification problem is a specific instantiation of the general framework, which exhibits the following additional properties:

P5. The cardinality of \mathcal{T} is m ; this implies that all targets are observed during training.

P6. No side information is available for targets. Again, without loss of generality, we can hence identify targets with natural numbers, such that the target space is $\mathcal{T} = \{1, \dots, m\}$.

P7. The score matrix Y has no missing values.

P8*. The score set is $\mathcal{Y} = \{0, 1\}$.

Definition 4 (Multi-task learning) A multi-task learning problem is a specific instantiation of the general framework, which exhibits the following additional properties:

P5. The cardinality of \mathcal{T} is m ; this implies that all targets are observed during training.

P6. No side information is available for targets. Again, the target space can hence be taken as $\mathcal{T} = \{1, \dots, m\}$.

P8**. The score set is homogenous across columns of Y , e.g., $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \mathbb{R}$.

In the above definitions, we adopt a matrix view for multi-task learning. This means we consider a matrix of targets, where rows are indexed by the number of training instances, and columns are indexed by the different tasks. A cell (i, j) is only filled with a value y_{ij} if instance i is contained in the training dataset of task j . We will use the letter Y to denote the resulting sparsely filled matrix. Figure 1-3 depicts an example of such a matrix.

From this matrix viewpoint, multivariate regression and multi-label classification arise as special cases (when the matrix Y has no missing values, like in the examples in Figure 1). Multi-task learning methods can usually process multivariate regression and multi-label classification datasets with no problems. In fact, many authors of multi-task learning papers turn out to be analyzing multivariate regression or multi-label classification datasets in their experimental studies, because such datasets are widely available. However, albeit not always visible in such experimental studies, multi-task learning methods are usually more general than multivariate regression and multi-label classification methods.

Multivariate regression and multi-label classification can be further generalized to targets with other types of values, such as nominal, ordinal, or mixed. This setting basically omits Property 8**. It is sometimes referred to as multi-dimensional classification, even though it has been rarely investigated in the machine learning literature—see e.g. (Bielza et al, 2011; Read, 2013). Another problem that can be seen as an instantiation of our framework is label ranking, where each instance is associated with a ranking (total order) of the targets (Hüllermeier et al, 2008). Thus, the score $y_{ij} \in \{1, \dots, m\}$ for a pair $(\mathbf{x}_i, \mathbf{t}_j)$ is the position of \mathbf{t} in the ranking associated with \mathbf{x} , i.e., each row of the score matrix Y is a permutation of $\{1, \dots, m\}$.

2.3 Problems that involve side information for targets

Let us now extend multivariate regression, multi-label classification, and multi-task learning to settings where additional side information about the target space is available. To this end, we adjust the three examples from above slightly.

Example 4 In the drug design application of Example 1, we treated proteins as instances and small molecules that can potentially bind to those proteins as targets. Let us assume that, in addition to the graph-based representation for the proteins, a representation for the target molecules is also available, as shown in Figure 1-4. This is a common situation in research on protein-ligand prediction, where the representation for the targets is either graph-based or feature-based. The resulting machine learning setting can be interpreted as a dyadic prediction problem.

Example 5 Let us return to the document categorization application that was described in Example 2. Often document categories are organized in a hierarchy that describes the degree of relatedness among document tags. An example is shown in Figure 1-5. Problems of that kind, where the target space is equipped with a hierarchical structure, are referred to as hierarchical multi-label classification problems. Naturally, state-of-the-art algorithms will try to exploit the structure for better prediction.

Example 6 In the student mark forecasting application that was considered before, one might also assume that additional information about the targets is given. One might be able to collect all sorts of variables about schools and courses, such as geographical location, qualifications of the teachers, reputation of the school, etc. Figure 1-6 visualizes such a situation with binary features. One might be able to improve multi-task learning algorithms by taking the features into account. This side information forms a key element for tackling transfer learning and zero-shot learning problems.

Examples 4-6 further extend basic MTP settings to situations where various types of side information about targets is available. Example 4 illustrates this with the case of a structured representation in the form of a molecular graph. Example 5 assumes a hierarchy that describes relations among targets in a specific manner, and Example 6 considers feature representations for targets. Applications of that kind lead to more complex multi-target prediction problems that are often referred to as dyadic prediction, link prediction, or network inference settings—see e.g. (Menon and Elkan, 2010; Schäfer and Hüllermeier, 2015). In this area, one can distinguish algorithms that model vector representations or structured target representations, as well as methods that model target relations. This will be further discussed in Section 3.

Dyadic prediction, link prediction, and network inference are general terms that cover a wide range of problems. Generally speaking, they cover problems that obey the four properties listed in Definition 1. The labels y_{ij} can be arranged in a matrix Y , which is often sparsely filled. Thus, one may argue that dyadic prediction is nothing else than multi-task learning with task features. However, the multi-task learning terminology is rarely used in the dyadic prediction literature. Dyadic prediction problems emerge in a variety of application domains, including product recommendation, social network analysis, drug design, various bioinformatics applications, and game playing (Basilico and Hofmann, 2004; Park and Chu, 2009; Stock et al, 2014; Ben-Hur and Noble, 2005; Kashima et al, 2009; Pelosof et al, 2015; Jacob and Vert, 2008; Pahikkala et al, 2010, 2013).

2.4 Inductive versus transductive learning problems

One may argue that the problems analyzed in Examples 1-3 are inductive w.r.t. instances and transductive w.r.t. targets. Predictions need be generated for novel instances, whereas the set of targets is known beforehand and observed during the training phase. Side information

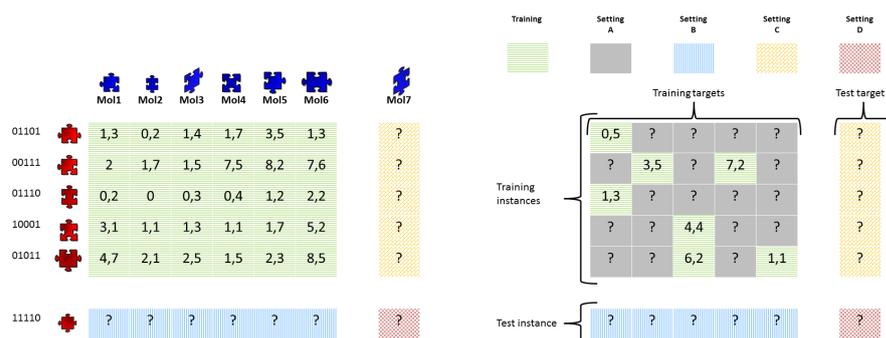


Fig. 2 Visualization of the different types of prediction problems that arise when prior knowledge of the target space is modelled. Left: an extension of the drug design example, where predictions have to be made for a novel protein (Setting B in blue), a new ligand (Setting C in yellow) or a combination of a new protein and ligand (Setting D in red). Right: a general overview of the four settings that can be distinguished. Setting A depicts the situation where missing values need to be imputed in the matrix. See text for details.

is of crucial importance for generalizing to novel targets that are unobserved during the training phase, such as a novel target molecule in the drug design example, a novel tag in the document annotation example, or a novel course in the student grading example.

This is visualized in Figure 2, where the red part of the dataset alludes to a protein-drug pair that was not observed during the training phase. In the multi-task and transfer learning literature, this type of problem setting is known as zero-shot learning (Larochelle et al, 2008). Since no training data is available for the novel target, the problem is intrinsically more difficult than those analyzed in Section 2.2 (Park and Marcotte, 2012). In this work, we define zero-shot learning as follows.

Definition 5 (Zero-shot learning) A zero-shot learning problem is a specific instantiation of the general framework with the following additional property:

P5*. The cardinality m^* of the set \mathcal{T} is bigger than m . Some targets are hence not observed during training, but may nevertheless appear at prediction time.

By substituting P5 with P5*, one now tackles problems that are inductive instead of transductive w.r.t. targets. The same subdivision can be made for instances. As pointed out by Pahikkala et al (2014), in total four different settings can be distinguished in the presence of side information about instances and targets. In Figure 2, Setting D resembles the zero-shot learning setting, whereas Setting B coincides with the problems that we discussed in Section 2.2. In addition, one can also think of Setting C (predicting scores for a novel target on an instance that was contained in the training dataset) and Setting A (predicting scores for an instance-target combination, where the target and the instance were already seen during the training phase). Combinations of those settings also exist. For example, one can distinguish zero-shot learning settings where only targets that are unknown during training need to be predicted in the test phase (thus Setting D), versus settings where both training and test targets have to be predicted for novel instances (a combination of Setting B and D). The latter, for example, happens in image classification tasks, where it is referred to as generalized zero-shot learning (Rohrbach et al, 2011; Xian et al, 2017). Setting C is inductive w.r.t. targets and transductive w.r.t. instances. Setting A is transductive w.r.t. instances and targets.

Setting C is in theory identical to Setting B. When side information is available for both instances and targets, the two settings become interchangeable. In Example 4, we referred

to proteins as instances and small molecules as targets, but in principle, one could transpose the matrix without changing the algorithms. However, interchanging the rows and columns of a particular dataset and applying a particular multi-target prediction algorithm on the two scenarios will in practice lead to different results. Often, one of the two sources of side information is more informative than the other, and many algorithms take this implicitly into account. Take the document categorization example, for which textual descriptors on the document level are much more informative and richer than the hierarchical structure on the tag level. Therefore, modelling the side information on the document level is of uttermost importance, while a reasonable performance might still be expected even if the side information on the tag level is ignored.

In contrast, Setting A results in yet a different type of problem, namely that of matrix completion. Both the targets and the instances are observed, albeit not for all instance-target combinations. In principle, one can solve this setting with methods that ignore side information of the target and the instance space. In such cases, latent representations of instance and target space are both deduced using only the instance-target interaction matrix. In light of Examples 1-3, this means that the feature representations of proteins, documents, and students are not used. In fact, they do not even have to be available. As a result, we formally define matrix completion problems as follows.

Definition 6 (Matrix completion) A matrix completion problem is a specific instantiation of the general framework with the following additional properties:

- P5. The cardinality of \mathcal{T} is m . This implies that all targets are observed during training.
- P6. No side information is available for targets. Without loss of generality, we can hence assign identifiers to targets from the set $\{1, \dots, m\}$ such that the target space is $\mathcal{T} = \{1, \dots, m\}$.
- P9. The cardinality of \mathcal{X} is n . This implies that all instances are observed during training.
- P10. No side information is available for instances. Without loss of generality, we can hence assign identifiers to instances from the set $\{1, \dots, n\}$, such that the instance space is $\mathcal{X} = \{1, \dots, n\}$.

Matrix completion methods are extensively used in areas such as recommender systems, social network analysis, and biological network inference. They are therefore also known as collaborative filtering and link prediction methods. One can distinguish different versions of matrix completion. One may consider matrices that contain continuous, ordinal or binary values, but situations with presence-only data are also common. Observed entries in the matrix then correspond to known positive interactions between the items in the rows and the columns, and the goal consists of predicting more positive interactions from the large pool of missing entries. The assumption made in such cases is that the overwhelming majority of the missing entries correspond to negative interactions, with the exception of a few positive ones that need to be found. Properties 5 and 9 characterize the transductiveness of matrix completion methods. Properties 6 and 10 characterize the difference between a pure matrix completion setting, and a hybrid setting that simultaneously exploits the structure of the matrix and side information to generate predictions.

2.5 Beyond MTP: Problems that are not covered

Our formal framework is rather generic and covers several machine learning problems as special cases, including those discussed above. In principle, every prediction problem with

(original) output space \mathcal{Y} could be seen as a special case, simply by taking $\mathcal{T} = \mathcal{Y}$ as the set of tasks and $\{0, 1\}$ as a score set. This amounts to treating each candidate output as a target, and reinterpreting the task to predict an output from \mathcal{Y} as predicting, for each candidate value, whether or not the sought output corresponds to that candidate. Consequently, a consistent prediction has to obey strong (deterministic) dependencies between the targets: It should assign the score 1 to exactly one target and 0 to all the others. As two examples, consider multi-class classification and structured output prediction (SOP), which can be formalized, respectively, as specializations of multi-label classification and zero-shot learning.

Definition 7 (Multi-class classification) A multi-class classification problem is a specific instantiation of the general framework with the following additional properties:

- P5. The cardinality of \mathcal{T} is m . This implies that all targets are observed during training.
- P6. No side information is available for targets. Without loss of generality, we can hence assign identifiers to targets from the set $\{1, \dots, m\}$ such that the target space is $\mathcal{T} = \{1, \dots, m\}$.
- P7. The score matrix Y has no missing values.
- P8*. The score set is $\mathcal{Y} = \{0, 1\}$.
- P11. Each row in Y has a single “positive” entry, and all other entries are zero.

Definition 8 (Structured output prediction) A structured output prediction problem is a specific instantiation of the general framework with the following additional properties:

- P5*. The cardinality m^* of the set \mathcal{T} is bigger than m . Some targets are hence not observed during training, but they appear at test time.
- P7. The score matrix Y has no missing values.
- P8*. The score set is $\mathcal{Y} = \{0, 1\}$.
- P11. Each row in Y has a single “positive” entry, and all other entries are zero.

Note that these encodings of the problems correspond to what is known as a 1-versus-rest decomposition, a specific reduction technique that makes prediction problems with $|\mathcal{Y}| > 2$ amenable to binary classification. Obviously, other reduction techniques such as all-pairs or, more generally, error correcting output codes (ECOC) lead to similar representations: The problem of predicting the output for an instance $\mathbf{x} \in \mathcal{X}$ is decomposed into a set of binary prediction problems, and each of these problems is seen as a task.

While formally possible, we are not in favor of considering problems such as multi-class classification and structured output prediction as special cases of MTP. The reasons include both conceptual and algorithmic aspects. Conceptually, the view of each candidate prediction as a separate target appears to be rather artificial. Actually, one is still interested in a single prediction, not multiple ones. To comply with the corresponding consistency constraints, a kind of post-processing (like the decoding step in ECOC) is normally required. Algorithmically, it is also clear that the multi-target perspective is not typical of methods for SOP. Instead, such methods are specifically tailored for output spaces in SOP, which are often huge but equipped with a strong structure.

In this paper, we will also exclude other (multi-target) prediction problems where the ground truth cannot be represented in a matrix format with optional side information for rows and columns. This includes problems that involve multi-instance learning representations (Zhou and Zhang, 2007) and dyadic feature representations (Van Peer et al, 2017). The latter scenario occurs, for example, in bio-informatics and recommender systems applications, when features are available that describe an interaction between an instance and a target, e.g., the day of the week when a user clicked a specific item on a website.

3 A unifying view on MTP methods

In this section, we present a unifying view on MTP methods. To this end, we categorize methods according to some of the properties that were discussed in Section 2. In particular, we believe that the properties characterizing Settings A, B, C and D are specifically important in this regard, and suitable to distinguish MTP methods. We therefore put a main emphasis on those properties, while some other properties of Section 2 are not further considered. As a result, our categorization of MTP methods does not yield a one-to-one mapping with MTP problems. Table 1 gives an overview of the subdivision we propose. Most of the methods to be discussed can be applied to Settings B and C. Some of them are also applicable to Setting D, while Setting A normally calls for more specialized methods.

Table 1 Categorization of MTP methods.

Group of methods	Applicable setting
Section 3.2: Similarity-enforcing methods	B and C
Section 3.3: Relation-exploiting methods	B, C and D
Section 3.4: Relation-constructing methods	B and C
Section 3.5: Representation-exploiting methods	B, C and D
Section 3.6: Representation-constructing methods	B and C
Section 3.7: Matrix completion and hybrid methods	A

In what follows, we focus on the average squared prediction error over all targets as a loss function to be minimized. Formally, for a given pair (\mathbf{x}, \mathbf{t}) , a predictor f yields a prediction $\hat{y} = f(\mathbf{x}, \mathbf{t})$ of an actual outcome (score) y . Let \mathbf{y} be a vector of actual output values of size m (or m^* in the case of zero-shot learning), and let $\hat{\mathbf{y}}$ be a vector of the corresponding predictions. Then, the squared-error loss is given by

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^m (y_i - \hat{y}_i)^2. \quad (1)$$

In spite of obvious shortcomings, both in regression and classification settings, this loss is somewhat representative in the sense of being used in the majority of algorithms for multivariate regression and multi-task learning, and therefore suitable for explaining general concepts. It has the additional advantage of being applicable for both numerical and binary target variables, again making it a natural choice for the unifying view that we intend to provide. All the methods presented below can be tailored for minimizing this loss function.

In multi-label classification, many methods optimize loss functions that are more complex and not decomposable over targets. Because those methods only appear in the MLC literature, an in-depth discussion is beyond the scope of this paper. Instead, we refer the interested reader to Dembczyński et al (2012).

We start our discussion on multi-target methods by introducing a simple baseline that does not exploit any target dependencies. Subsequently, we present more and more complex models that tackle the problem of multi-target prediction in different manners.

3.1 Independent models

The most straight-forward approach to solving MTP problems consists of constructing one model f_i for every target independently, and to concatenate the predictions of these models into the sought multi-target prediction. In the multi-label classification community, this approach has an own name and is known as *binary relevance learning* (Tsoumakas and Katakis, 2007). In order to introduce the approach more formally, let $f_i(\mathbf{x})$ denote the score (continuous or discrete) assigned to the i th target for instance \mathbf{x} . Using linear basis function models, the model for the i th target can be represented as

$$f_i(\mathbf{x}) = \mathbf{a}_i^\top \phi(\mathbf{x}), \quad (2)$$

where \mathbf{a}_i is a target-specific parameter vector and $\phi(\mathbf{x})$ a feature representation that is either given explicitly or first obtained from another representation in a preprocessing step. Multivariate ridge regression is probably the most basic model of this kind. This approach fits a linear model in a regularized least-squares fashion, while ignoring potential dependencies among targets (Hastie et al, 2007). More formally, multivariate ridge regression boils down to minimizing the following objective function:

$$\min_A \|Y - XA\|_F^2 + \sum_{i=1}^m \lambda_i \|\mathbf{a}_i\|^2, \quad (3)$$

where Y is the $n \times m$ target matrix with n the number of observations and m the number of targets. Denoting the number of features by p , X and A are $n \times p$ and $p \times m$ matrices, respectively, which are constructed as follows:

$$X = \begin{bmatrix} \phi(\mathbf{x}_1)^T \\ \vdots \\ \phi(\mathbf{x}_n)^T \end{bmatrix} \quad A = [\mathbf{a}_1 \quad \cdots \quad \mathbf{a}_m]. \quad (4)$$

$\|\cdot\|_F$ is the Frobenius-norm, so that the left-hand side corresponds to the sum of squared error losses (1) on the m targets. The λ_i are regularization parameters.

Optimization problem (2) can be solved by matrix inversion in a least-squares algebraic fashion. This is efficient because the inversion needs to be done only once for all targets. The formulation assumes a linear statistical model, but independent models can be constructed using any well-known method for regression (in the case of continuous variables as targets) or classification (in the case of binary variables as targets).

Can one perform better with more sophisticated approaches that exploit dependencies among targets? Unsurprisingly, the answer is affirmative. Many of the approaches that were introduced over the last decade in each of the MTP subcommunities are able to outperform the simple baseline that treats every target as an independent problem. Generally speaking, these methods put soft constraints on the values that targets can take, so they assume that not all combinations of values are equally likely to occur: some combinations of target values have a higher probability of occurrence than others, yielding a dependency among targets in a statistical sense.

Authors introducing a novel multi-target prediction method often emphasize the exploitation of target dependencies as the key to improved performance. Without questioning explanations of that kind, we like to note that improvements over independent predictions can be achieved even for problems without any statistical dependence between the targets. This insight is illustrated, for example, by the famous James-Stein paradox, showing that,

under certain conditions, maximum likelihood estimation is not optimal (or, using statistical terminology, not admissible) for estimating the mean of a multivariate Gaussian (James and Stein, 1961). Since the James-Stein estimation principle is less known among machine learning scholars, we provide a short description in Appendix A.

3.2 Similarity-enforcing methods

We start our overview of MTP methods by explaining a few simple methods that do not consider any particular domain knowledge about targets. Such methods can be found in multivariate regression, multi-label classification, and multi-task learning, and they are usually only applicable to Settings B and C. The central idea in those simple methods is that models for different targets should behave *similar*. This can be achieved, for example, by enforcing similarity of the parameterizations of models for different targets. To this end, one usually defines a mathematical objective that incorporates a specific regularizer. In early work on multi-task learning, Evgeniou and Pontil (2004), Evgeniou (2005), and Jalali et al (2010) solve the following optimization problem:

$$\min_A \|Y - XA\|_F^2 + \lambda \sum_{i=1}^m \|\mathbf{a}_i - \frac{1}{m} \sum_{j=1}^m \mathbf{a}_j\|^2, \quad (5)$$

with A the parameter matrix, X the feature matrix, and Y the target matrix, as defined before. In this way, one penalizes for deviations from the “mean target”. The parameter estimates are biased to the mean, in an attempt to reduce the variance and prevent overfitting.

Another common way of obtaining restricted models with less flexibility (compared to modelling each target independently) consists of implementing feature selection strategies that retain the same features for different targets. Joint feature selection leads to a model with less parameters compared to binary relevance, and this may in turn result in performance gains. The idea has been particularly popular in the area of multi-label classification (Obozinski et al, 2010; Gu et al, 2011; Kong and Yu, 2012; Spolaôr et al, 2016). Even though most of those authors put the emphasis on computational efficiency as a motivation, selecting a subset of identical features for different targets can also improve the predictive performance. Similar ideas are appearing in the literature on multi-task learning (Zhou et al, 2011b) and in joint regression of binary and continuous response variables (Zhang and Shen, 2012). An embedded feature selection strategy can be obtained, for example, by a combination of L1 and L2 norms:

$$\min_A \|Y - XA\|_F^2 + \lambda \sum_{j=1}^p \|\mathbf{a}_j\|^2$$

In contrast to (5), the sum is taken over the features. Unlike (4), the vectors \mathbf{a}_j now represent the rows of matrix A :

$$A = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{bmatrix}$$

Thus, \mathbf{a}_j specifies the set of parameters for the j th feature, across different targets. The combination of the two types of norms results in group sparsity, so that a set of identical features is selected for different targets.

Yet another straight-forward approach to exploiting dependencies among targets is stacking (Wolpert, 1992). This approach was initially introduced as an ensemble learning technique in conventional classification and regression settings. Later on, it has been extended for multi-label classification (Godbole and Sarawagi, 2004; Cheng and Hüllermeier, 2009). In the statistical literature, a slightly more advanced method has shown to improve the predictions for multivariate regression (Breiman and Friedman, 1997). Although one could easily think of extensions for multi-task learning problems, too, the method has been less popular in that area.

Stacking implements a two-step procedure. In the first step, a model is fitted for every target individually, as discussed in Section 3.1. Then, in a second step, the predictions obtained by each model are used as a feature representation to train a second series of models, again one for each target. Obviously, these second-order models are able to capture dependencies between different targets, as they seek to represent the prediction of one target as a function of the (predictions of) the others. Thus, the goal of the second step is to discover dependencies among targets, and to identify, for each target in turn, a set of other targets that help improve predictions. More formally, with the shorthand-notation

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix}$$

for the independent models of (2), the general scheme of stacking can be expressed as follows:

$$\mathbf{f}^*(\mathbf{x}) = \mathbf{b}(\mathbf{f}(\mathbf{x}), \mathbf{x}), \quad (6)$$

with $\mathbf{f}^*(\cdot)$ the predictions made by stacking and $\mathbf{b}(\cdot)$ a second-stage model that shrinks or regularizes the solution of the initial models. Breiman and Friedman (1997) show formally that stacking prevents overfitting. As soon as some regularization is present in the second-order models, one introduces a bias by encouraging models for different targets to learn similar parameters. The same authors also reveal some connections with James-Stein estimation.

Two appealing properties of stacking are its ease of implementation and its generality, as it can be applied in tandem with any type of classifier or regressor. In addition, the second-level model \mathbf{b} can be trained on the first-level predictions $\mathbf{h}(\mathbf{x})$ alone or in concatenation with the original features \mathbf{x} . In the case of classification, there is also a choice between feeding the second-level model with binary predictions or with continuous scores obtained from scoring functions or probabilities, if such outputs are delivered by the classifier.

The idea of enforcing that models for different targets behave similarly can also be found in neural networks, starting from classical papers such as (Caruana, 1997), till modern deep learning architectures, such as deep convolutional nets for multi-label classification (Wei et al, 2016). Such methods are based on weight sharing in the final network layer, which usually includes m nodes for a multi-target prediction setting with m targets. Likewise, the common idea of pre-training a deep neural network on an auxiliary task with a lot of training data, and further optimizing it for another task with limited training data, can be interpreted as weight sharing—see e.g. (Donahue et al, 2014; Girshick et al, 2014; Oquab et al, 2014; Razavian et al, 2014; Sermanet et al, 2014; Gong et al, 2014).

3.3 Relation-exploiting methods

In this section, we discuss MTP methods using side information in the form of target relations, such as hierarchies, graph structures, decision rules, and correlation matrices. This side information might be useful for at least three purposes. First, it might help obtain performance gains for Settings B and C, compared to the methods that were discussed in Section 3.2. Second, it can be considered as a key element when generalizing to novel targets in Setting D. Third, it may also effect the computational scalability of algorithms in a positive way.

When speaking about target relations, it is important to make a distinction between deterministic relations, which are guaranteed to hold for each observation, and probabilistic relations, which only give an indication of the likelihood of occurrence. In multi-label classification, for example, one could observe deterministic relations such as implications, subsumptions, and mutual exclusions between relevant labels of an instance. In label ranking, the values of target variables are positions in a ranking, which means that all target values must be different (each position can only be occupied once). In multivariate regression, targets may represent consecutive time steps, for which constraints such as smoothness or monotonicity over consecutive values might apply.

For MLC problems such as Example 5 in Section 2, relationships between labels are often represented in the form of a tree or hierarchy, resulting in a set of problems that is known as hierarchical multi-label classification. Many authors have shown that such a hierarchy can be integrated in specific methods (Rousu et al, 2006; Barutcuoglu et al, 2006; Vens et al, 2008; Silla and Freitas, 2010; Gopal et al, 2012; Nam et al, 2015). The hierarchy not only helps to improve predictions, but also allows for defining extended prediction settings that are characterized by specific performance measures. For example, when a classifier is unsure about certain classes, it could be allowed to return intermediate nodes from the tree as prediction instead of the leaf nodes that correspond to single labels (Bi and Kwok, 2012).

A few authors have presented more general approaches for incorporating deterministic relations that cannot be represented as a hierarchy. Gopal and Yang (2013) and Deng et al (2014) both consider graph structures. When targets follow a chronological order, chain-based graph structures might be a natural choice. For example, Zhou et al (2012a) study disease progression prediction as a multi-task learning problem, where tasks follow a chronological order, leading to a formulation where a specific temporal regularizer in the optimization problem is introduced. This regularizer can be seen as a special case of the graph-based regularization of Gopal and Yang (2013):

$$\min_A \|Y - XA\|_F^2 + \lambda \sum_{i=1}^m \sum_{j \in \mathcal{N}(i)} \|\mathbf{a}_i - \mathbf{a}_j\|^2,$$

where $\mathcal{N}(i)$ denotes the set of targets that are related to the i th target. This formulation shows that methods using target relations are actually very related to those exploiting similarity (cf. Section 3.2). The available side information essentially allows one to define prior knowledge concerning statistical dependencies that are likely to occur.

In addition to deterministic relations, there are also methods that incorporate probabilistic relations. In the context of kernel methods, it is natural to represent target similarities by means of so-called output kernels, which can be interpreted as expressing some sort of correlation between targets. Output kernels have been mainly developed for multivariate regression, under the framework of vector-valued kernel functions (Caponnetto et al, 2008;

Baldassarre et al, 2012; Álvarez et al, 2012). Similar kernel-based formulations have been proposed for multi-label classification (Hariharan et al, 2010).

All the above methods essentially assume that the provided information about the target space results in a good representation for modelling target dependencies. This is clearly a necessary condition for improvement—a few authors have also reported performance drops in certain applications, when the given target relations lead to an incorrect representation of target dependencies.

3.4 Relation-constructing methods

When target relations to be used as side information are not available in the form of prior knowledge, one can try to construct such relations from the training data. As an important consequence of this approach, note that it excludes a generalization to novel targets. Thus, unlike the methods discussed in the previous section, methods that learn target relations are not applicable to Setting D; instead, they are mainly useful for settings B and C.

Learning target relations can be useful for several reasons. First of all, domain knowledge is often not available, so that target relations cannot be modelled as prior information. Second, even if such knowledge can in principle be provided, target relations specified by domain experts might be incorrect and hence misleading for machine learning algorithms. A third motivation is computational efficiency, since learning target relations paves the way for methods that scale sublinearly in the number of targets.

Like in the case where target relations are given instead of being learned, one can distinguish methods w.r.t. the type of relations: hierarchies, more general graph structures, and correlation matrices. In hierarchical multi-label classification, where the correctness of a hierarchy could be questioned, one can find several methods that adjust the existing hierarchy in the course of the learning procedure. In that sense, one can distinguish level-flattening algorithms, node removal algorithms, hierarchy modification algorithms, and hierarchy generation algorithms; see (Rangwala and Naik, 2017) for an overview. In those methods, the focus is on improving the predictive performance. Conversely, in extreme multi-label classification, i.e., problems that involve thousands or millions of labels, hierarchies are often learned to obtain a training or prediction time that is logarithmic in the number of labels (Agrawal et al, 2013; Weston et al, 2013; Prabhu and Varma, 2014; Dembczynski et al, 2016).

Graphs more general than hierarchies might be considered. This idea is popular when applying graphical models in multi-target prediction settings, see e.g. (Guo and Gu, 2011; Papagiannopoulou et al, 2015). Using directed graphs, one can infer asymmetric relationships among targets (Lee et al, 2016), while similar results might be obtained with specific rule-based systems (Park and Fürnkranz, 2008; Loza Mencía and Janssen, 2016). Let us remark that, for many of those methods, it becomes more difficult to say whether the inferred target relations are deterministic or probabilistic, because the learning process usually contains both deterministic and probabilistic elements.

In some applications, one may cluster targets into a number of non-overlapping groups, such that interactions between targets need to be considered only inside the same group. This can be established by using a cluster-norm in the optimization (Jacob et al, 2008; Wang et al, 2009). Similarly, clustered multi-task learning was considered in a Bayesian setting by Bakker and Heskes (2003) via a mixture of Gaussians instead of a single Gaussian prior, and by Xue et al (2007) via a Dirichlet process prior. Furthermore, Gong et al (2012) identify outlier targets, resulting in a structure where only information is shared among the inliers.

Similarly, clustering methods have been considered in multi-label classification, where both performance gains in prediction and computational complexity have been reported.

Finally, one can also learn target relations in the form of similarity or correlation matrices. Unsurprisingly, this idea has been very popular in the field of kernel methods. The older methods rather assume that the output kernel is given as domain knowledge, whereas more recent methods try to discover target similarities by learning the output kernel in a convex formulation (Dinuzzo et al, 2011; Dinuzzo, 2013; Jawanpuria et al, 2015). Similarly, one can learn the covariance function among targets in a Bayesian treatment, by assuming a multivariate Gaussian distribution as prior (Zhang and Yeung, 2010).

3.5 Representation-exploiting methods

Instead of using relations between targets, one can also consider to use target representations as side information. Examples 4 and 6 in Section 2 describe two applications where this type of side information is available in the form of a graph representation and a vector representation, respectively. Just like for target relations, this side information can be useful for several purposes. It might boost the predictive performance in Settings B and C, and is essential for generalizing to novel targets in Setting D.

Compared to target relations, target representations still require the learning algorithm to discover which elements of the representations are useful and which are not. For instance, in the student grading application of Example 6, one still has to figure out which of the school features are useful. Likewise, in the drug design application of Example 4, one would need specialized methods that can handle structured data such as molecular graphs. In contrast, when working with target relations, the side information is usually available in a more direct and a less noisy format.

Historically, kernel methods have played an important role in modelling target representations. Such methods can easily process vectorial representations, as well as structured data. Moreover, they also establish a clear connection with methods that model target relations: a transformation from a primal formulation to a dual formulation in fact implies a transformation from a target representation to a target relation. In the primal formulation, information about the instance space and the target space is combined by means of a joint feature representation, whereas in the dual a joint kernel is modelled. Such representations typically yield prediction models of the following form:

$$f(\mathbf{x}, \mathbf{t}) = \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{t}) = \sum_{(\bar{\mathbf{x}}, \bar{\mathbf{t}}) \in \mathcal{D}} \alpha_{(\bar{\mathbf{x}}, \bar{\mathbf{t}})} \Gamma((\mathbf{x}, \mathbf{t}), (\bar{\mathbf{x}}, \bar{\mathbf{t}})), \quad (7)$$

where \mathcal{D} represents the training data as defined before, \mathbf{w} and $\alpha_{(\bar{\mathbf{x}}, \bar{\mathbf{t}})}$ are primal and dual parameters, $\Psi(\mathbf{x}, \mathbf{t})$ is a joint feature representation in the primal formulation, and

$$\Gamma((\mathbf{x}, \mathbf{t}), (\bar{\mathbf{x}}, \bar{\mathbf{t}})) = \Psi(\mathbf{x}, \mathbf{t})^T \Psi(\bar{\mathbf{x}}, \bar{\mathbf{t}})$$

is a joint kernel in the dual formulation, with \mathbf{t} and $\bar{\mathbf{t}}$ representations of two targets.

The most commonly used pairwise kernel is the Kronecker product pairwise kernel (Basilico and Hofmann, 2004; Oyama and Manning, 2004; Ben-Hur and Noble, 2005; Park and Chu, 2009; Hayashi et al, 2012; Bonilla et al, 2007; Pahikkala et al, 2013), resulting in the following mathematical model:

$$\Psi(\mathbf{x}, \mathbf{t}) = \phi(\mathbf{x}) \otimes \psi(\mathbf{t}), \quad \Gamma((\mathbf{x}, \mathbf{t}), (\bar{\mathbf{x}}, \bar{\mathbf{t}})) = k(\mathbf{x}, \bar{\mathbf{x}}) \cdot g(\mathbf{t}, \bar{\mathbf{t}}),$$

with ϕ and ψ feature maps in primal form, and $k(\mathbf{x}, \bar{\mathbf{x}})$ and $g(\mathbf{t}, \bar{\mathbf{t}})$ any kernel on the instance and target space in dual form. For example, for the drug design application of Example 4, one would typically develop domain-specific kernels for both $k(\mathbf{x}, \bar{\mathbf{x}})$ and $g(\mathbf{t}, \bar{\mathbf{t}})$. In the student grading application of Example 6, one would more likely apply conventional kernels for vectors on the given feature representations. For the document categorization application of Example 5, one would perhaps prefer to use other methods from the domain of hierarchical multi-label classification, but one could also apply a kernel that mimics the hierarchical structure, such as the shortest path kernel.

As an aggregation of the instance kernel $k(\mathbf{x}, \bar{\mathbf{x}})$ and the target kernel $g(\mathbf{t}, \bar{\mathbf{t}})$, the Kronecker product pairwise kernel might be a reasonable first choice in experimental studies, as it is known to exhibit universal approximation properties (Stock et al, 2016). Variants also exist for modelling additional domain knowledge concerning relationships between instances and targets (Vert et al, 2007; Pahikkala et al, 2010; Waegeman et al, 2012; Pahikkala et al, 2013)). Pairwise kernels are much more general building blocks than the output kernels that were discussed in Sections 3.3 and 3.4. Pairwise kernels model a similarity score based on target representations, whereas no target representations are available for output kernels.

Many pairwise kernels can be easily plugged into standard kernel methods for binary classification or regression, such as support vector machines or kernel ridge regression. Approaches of that kind are often criticized for computational reasons, and claimed to be infeasible for large sample studies, but efficient implementations based on algebraic shortcuts exist (Van Loan, 2000; Kashima et al, 2009; Raymond and Kashima, 2010; Álvarez et al, 2012; Pahikkala et al, 2013). More recently, Pahikkala et al (2014) and Romera-Paredes and Torr (2015) independently proposed algorithms that avoid the explicit construction of pairwise kernels. Those algorithms have computational advantages w.r.t. cross-validation and online training.

Most importantly, pairwise learning methods are capable of generalizing to zero-shot problems, such as a novel target molecule in the drug design example, a novel tag in the document annotation example, or a novel course in the student grading example. In recent years, specific zero-shot learning methods based on deep learning have become extremely popular in image classification applications. The central idea in all those methods is to construct semantic feature representations for class labels, for which various techniques might work. One class of methods constructs binary vectors of visual attributes (Lampert et al, 2009; Palatucci et al, 2009; Liu et al, 2011; Fu et al, 2013). Figure 3 shows an example of two such vectors for two images. Another class of methods rather considers continuous word vectors that describe linguistic context of images (Mikolov et al, 2013; Frome et al, 2013; Socher et al, 2013). For the same two images of Figure 3, one could for example look up “zebra” and “whale” on Wikipedia, and extract a Word2Vec representation from this source.

Many zero-shot learning methods for image classification adopt principles that originate in kernel methods. The model structure can often be formalized as follows:

$$f(\mathbf{x}, \mathbf{t}) = \mathbf{w}^T (\phi(\mathbf{x}) \otimes \psi(\mathbf{t})) \quad (8)$$

This model in fact coincides with the primal formulation of (8) with a Kronecker-based choice for $\Psi(\mathbf{x}, \mathbf{t})$. Different optimization problems with this model have been proposed (Frome et al, 2013; Akata et al, 2015, 2016), and related methods provide nonlinear extensions (Socher et al, 2013; Xian et al, 2016). Most of these optimization problems do not minimize squared error loss, and according to the definitions of Section 2, they should rather be seen as structured output prediction methods. Indeed, a representation such as (7) is in fact commonly used in structured output prediction. SOP methods additionally have inference procedures that allow for finding the best-scoring targets in an efficient manner.

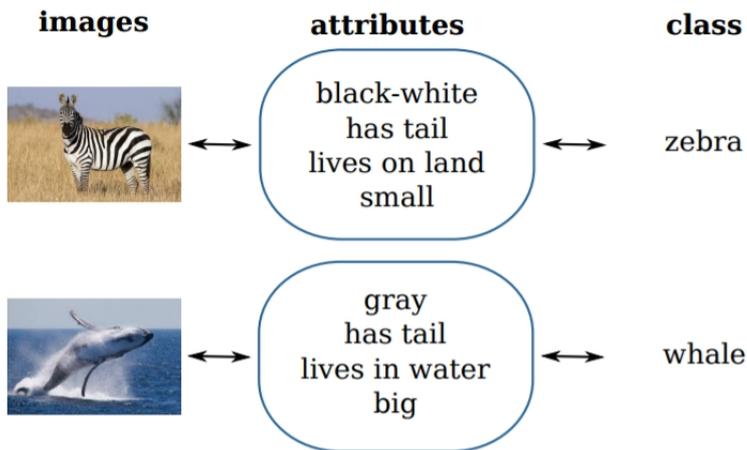


Fig. 3 Two examples of labeled images where the labels “zebra” and “whale” are described by four visual attributes. Examples taken from the CVPR 2016 Tutorial on Zero-shot learning for Computer Vision.

Some of the zero-shot learning methods from computer vision also turn out to be useful for the related field of text classification. For documents, it is natural to model a latent representation for both the (document) instances and class labels in a joint space (Nam et al, 2016).

3.6 Representation-constructing methods

In this section, we discuss methods that construct vector representations for targets without assuming side information as prior knowledge. So, in contrast to the methods that we discussed in Section 3.5, no target representations need to be available. Instead, such representations are learned from scratch in the training phase¹. As a result, these methods are mainly useful for Settings B and C, while not being applicable to Setting D. In principle, target representations can also be learned for Setting A. However, since this involves additional complications, methods for Setting A will be discussed in Section 3.7.

Methods that learn vector representations from scratch are in fact embedding-based methods. They all proceed from the assumption that the targets can be embedded in a vector space with a dimension lower than m . Principal component analysis (PCA), canonical correlation analysis (CCA), and partial least-squares (PLS) are simple techniques of that kind. In the case of PCA, one first embeds the targets in a low-dimensional space that is spanned by the most important principal components. In a second stage, one constructs regression models with the original features and the retained principal components as response variables. Predictions for new observations are then obtained by feeding them to the regression models and applying the inverse transformation to the original space. CCA and PLS both

¹ Let us remark that our notion of “representation-constructing” differs substantially from the notion of “representation learning” as commonly used in the area of deep neural networks. Here, we consider the construction of vector representations for targets. This is something that is not commonly done in multi-target prediction extensions of deep architectures.

solve a different formulation, in which targets and features are embedded jointly in a lower-dimensional space (Shawe-Taylor and Cristianini, 2004; Yu et al, 2006).

One can also obtain an embedding in another way, by introducing a rank constraint on the parameter matrix A during optimization:

$$\min_A \|Y - XA\|_F^2 + \lambda \text{rank}(A)$$

For example, Ando and Zhang (2005) and Chen et al (2009) have shown that such an approach yields performance improvements in multi-task learning. Similar formulations for multivariate regression can be found in the statistical literature, starting in the seventies with well-known methods such as reduced-rank regression (RRR) (Izenman, 1975) and FICYREG (Van der Merwe and Zidek, 1980). According to Breiman and Friedman (1997), these methods have the same generic form:

$$\mathbf{f}^*(\mathbf{x}) = (T^{-1}GT)A\mathbf{x},$$

where T is the matrix of sample canonical coordinates, the solution of canonical correlation analysis (CCA), and the diagonal matrix G contains the shrinkage factors for scaling the solutions of ordinary linear regression A . Based on this equation, Breiman and Friedman (1997) also establish a close connection with stacking, which was discussed in Section 3.2.

While embedding-based methods originated in multivariate regression, they have been used extensively for multi-label classification (as well as structured output prediction) in recent years. Nowadays they are very popular in the area of extreme multi-label classification, when the number of labels m is extremely large. Then, embedding-based methods are needed for computational reasons, namely to obtain reasonable training and prediction times. They mainly differ in the choice of compression or reduction technique used. One can distinguish methods based on PCA (Weston et al, 2002), CCA (Rai and Daumé III, 2009), compressed sensing (Hsu et al, 2009), singular value decomposition (Tai and Lin, 2010, 2012), output codes (Zhang and Schneider, 2011), landmark labels (Balasubramanian and Lebanon, 2012; Bi and Kwok, 2013), Bloom filters (Cissé et al, 2013), auto-encoders (Wicker et al, 2016), etc. In spite of their success, there is also evidence that embedding-based methods may have important shortcomings. In many extreme multi-label classification datasets, the low-rank assumption is violated as a result of the presence of many so-called tail labels. These are labels that appear very rarely in the dataset, sometimes less than five times in total, and recovering them with low-rank matrix approximation methods is very difficult (Bhatia et al, 2015; Yen et al, 2016).

3.7 Matrix completion and hybrid methods

In this section, we review matrix completion methods. In Section 2, such methods were claimed to be useful for an MTP setting with partially-observed matrices Y —in Figure 2 referred to as Setting A. Both the targets and the instances are observed, but not for all instance-target combinations. In Setting A, side information about instances or targets is not required per se. We hence distinguish between methods that ignore side information and methods that also exploit such information, in addition to analyzing the matrix Y .

Inspired by the Netflix challenge in 2006, the former type of methods has been mainly popular in the area of recommender systems. Those methods often impute missing values by computing a low-rank approximation of the sparsely-filled matrix Y , and many variants exist in the literature, including algorithms based on nuclear norm minimization (Candes

and Recht, 2008), Gaussian processes (Lawrence and Urtasun, 2009), probabilistic methods (Shan and Banerjee, 2010), spectral regularization (Mazumder et al, 2010), non-negative matrix factorization (Gaujoux and Seoighe, 2010), and alternating least-squares minimization (Jain et al, 2013). In addition to recommender systems, matrix factorization methods are commonly applied to social network analysis (Menon and Elkan, 2010), biological network inference (Gönen, 2012; Liu et al, 2015), and travel time estimation in car navigation systems (Dembczyński et al, 2013). Let us stress that those methods differ substantially from the methods that were discussed in Section 3.6. Here, matrix factorization is used to impute missing values in Y , while in Section 3.6 it was needed to reduce the dimensionality of the target space. In other words, here we are discussing methods that have been proposed for Setting A, whereas the methods from Section 3.6 are applicable to Setting B.

In addition to matrix factorization, a few other methods exist for Setting A. Historically, memory-based collaborative filtering has been popular, and corresponding methods are very easy to implement. They make predictions for the unknown cells of the matrix by modelling a similarity measure between either rows or columns—see e.g. (Takács et al, 2008). For example, when rows and columns correspond to users and items, respectively, then one can predict novel items for a particular user by searching for other users with similar interests. In the resulting nearest neighbor search, the cosine similarity often outperforms other similarity measures.

Many variants of matrix factorization and other collaborative methods have been presented, in which side information of rows and columns is considered during learning, in addition to exploiting the structure of the matrix Y —see e.g. (Basilico and Hofmann, 2004; Abernethy et al, 2008; Adams et al, 2010; Fang and Si, 2011; Zhou et al, 2011a; Menon and Elkan, 2011; Zhou et al, 2012b). One simple but effective method is to extract latent feature representations for instances and targets in a first step, and combine those latent features with explicit features in a second step, using any of the more conventional approaches that were discussed in the previous paragraphs (Volkovs and Zemel, 2012). In light of Examples 1–3, when feature representations are available for proteins, documents, and students, it would be pointless to ignore them. Hybrid methods seek to combine the best of both worlds, by simultaneously modeling side information and the structure of Y . In addition to Setting A, they can often be applied to Settings B and C, which coincide, respectively, with a novel user and a novel item in recommender systems. In that context, one often speaks about cold-start recommendations.

4 Conclusion and future perspectives

In this paper, we provided a unifying view on MTP problems and methods. In Section 2, we presented a general MTP framework and explained how well-known subfields of machine learning can be seen as specific instantiations of this framework. In Section 3, we gave an overview of MTP methods, categorized according to the prediction setting of interest and the domain knowledge that is available. For researchers who are new to the field of multi-target prediction, this categorization might help identify which method to use for a given MTP problem. As an overall conclusion, it should be clear that no MTP method is applicable, let alone optimal, under all conditions. It is up to the data scientist to choose the right method for a given problem, and we hope that this article will be helpful in this regard.

There are several topics related to multi-target prediction that are not discussed in this paper. One important issue is the loss one seeks to optimize. As briefly discussed in Section 3, we would like to emphasize that the methods reviewed in this paper are mainly use-

ful for optimizing loss functions that are decomposable over targets—our example of mean squared error is exactly of that kind. For more complex loss functions, which mainly seem to appear in the multi-label classification literature, other algorithms will be needed. We refer to our previous work for an overview on this topic, with a particular focus on the subset zero-one loss, the rank loss, and the F1-measure (Dembczyński et al, 2012; Waegeman et al, 2014).

We believe that research on multi-target prediction will be further intensified in the coming years, because novel applications are constantly emerging. Besides, several research questions still remain unanswered. In general, one can observe a trend towards analyzing larger target spaces, which require sophisticated algorithms that scale linearly or sublinearly in the number of targets. One also observes a trend towards extending classical machine learning paradigms for an MTP context—including semi-supervised learning, multi-instance learning, time series classification, data stream mining, network analysis, etc. We are convinced that extensions of that kind will be a driving force of fundamental machine learning research in the coming years.

Finally, there is also a need for further theoretical research in multi-target prediction. Many of the recent algorithms rely on a purely heuristic or intuitive motivation, which is then tested (and usually confirmed) in empirical studies on classical benchmark datasets. Most of the time, however, a deeper understanding of those algorithms and a theoretical explanation of their behavior are missing. Needless to say, an understanding of that kind is needed to mature the field of multi-target prediction, and to unify the different branches discussed in this paper.

Appendix A: James-Stein Estimation

In the late sixties, James and Stein discovered that the best estimator of the mean of a multivariate Gaussian distribution is not necessarily the maximum likelihood estimator. More formally, assume that θ is the unknown mean of a multivariate Gaussian distribution with dimension $m > 2$ and a diagonal covariance matrix. Consider a single observation \mathbf{y} randomly drawn from that distribution:

$$\mathbf{y} \sim \mathcal{N}(\theta, \sigma^2 I).$$

Using only this observation, the maximum-likelihood estimator for θ would be $\hat{\theta}_{ML} = \mathbf{y}$. James and Stein discovered that the maximum likelihood estimator is suboptimal in terms of mean squared error

$$\mathbb{E}[||\theta - \hat{\theta}|^2],$$

where the expectation is over the distribution of \mathbf{y} . (In general, the expectation is taken over all samples that contain a single observation \mathbf{y} . Later on we will shortly discuss a situation in which we draw more than one observation to compute the value of the estimator). An estimator with lower squared error can be obtained by applying a regularizer to the maximum likelihood estimator. In case σ^2 is known, the James-Stein estimator is defined as follows:

$$\hat{\theta}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{||\mathbf{y}||^2}\right) \mathbf{y}.$$

From a machine learning perspective, a regularizer is introduced that shrinks the estimate towards the zero vector, and hence reduces variance at the cost of introducing a bias. It has been shown that this biased estimator outperforms the maximum likelihood estimator

in terms of mean squared error. The result even holds when the covariance matrix is non-diagonal, but in view of the discussion concerning target dependence, it is most remarkable for diagonal covariance matrices. In fact, in the latter case, it means that joint target regularization will be beneficial even if targets are intrinsically independent. This is somewhat in contradiction with what is commonly assumed in the machine learning literature.

Let us notice, however, that the advantage of the James-Stein estimate over the maximum likelihood estimate will vanish for larger samples (of more than one observation). In the second term in parentheses, σ^2 is then divided by the size of the sample, so that the James-Stein estimate converges to the maximum likelihood estimate when the sample size grows to infinity.

The James-Stein paradox analyzes a very simple estimation setting, for which suboptimality of the maximum likelihood estimator can be proved analytically, but the principle extends to various multi-target prediction settings. By interpreting each component of θ as an individual target (and omitting the instance space, or reducing it to a single point), the maximum likelihood estimator coincides with independent model fitting, whereas the James-Stein estimator adopts a regularization mechanism that is very similar to most of the regularization techniques used in the machine learning literature. For some specific multivariate regression models, connections of that kind have been discussed in the statistical literature (Breiman and Friedman, 1997). As long as mean squared error is considered as a loss function and errors follow a Gaussian distribution, one can immediately extend the James-Stein paradox to multivariate regression settings by assuming that target vectors \mathbf{y} are generated according to the following statistical model:

$$\mathbf{y} \sim \mathcal{N}(\theta(\mathbf{x}), \sigma^2 I),$$

where the mean is now conditioned on the input space. For other loss functions, we are not aware of any formal analysis of that kind, but it might be expected that similar conclusions can be drawn.

References

- Abernethy J, Bach F, Evgeniou T, Vert JP (2008) A new approach to collaborative filtering: operator estimation with spectral regularization. *Journal of Machine Learning Research* 10:803–826
- Adams RP, Dahl GE, Murray I (2010) Incorporating side information into probabilistic matrix factorization using Gaussian processes. In: Grünwald P, Spirtes P (eds) *The 26th Conference on Uncertainty in Artificial Intelligence*, pp 1–9
- Agrawal R, Gupta A, Prabhu Y, Varma M (2013) Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In: *Proceedings of the 22Nd International Conference on World Wide Web, ACM, New York, NY, USA*, pp 13–24
- Akata Z, Reed SE, Walter D, Lee H, Schiele B (2015) Evaluation of output embeddings for fine-grained image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp 2927–2936
- Akata Z, Perronnin F, Harchaoui Z, Schmid C (2016) Label-embedding for image classification. *IEEE Trans Pattern Anal Mach Intell* 38(7):1425–1438
- Álvarez M, Rosasco L, Lawrence N (2012) Kernels for vector-valued functions: a review. *Foundation and Trends in Machine Learning* 4(3):195–266
- Ando RK, Zhang T (2005) A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6:1817–1853

- Bakker B, Heskes T (2003) Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research* 4:83–99
- Balasubramanian K, Lebanon G (2012) The landmark selection method for multiple output prediction. In: *International Conference on Machine Learning*
- Baldassarre L, Rosasco L, Barla A, Verri A (2012) Multi-output learning via spectral filtering. *Machine Learning* 87(3):259–301
- Barutcuoglu Z, Schapire RE, Troyanskaya OG (2006) Hierarchical multi-label prediction of gene function. *Bioinformatics* 22(7):830–836
- Basilico J, Hofmann T (2004) Unifying collaborative and content-based filtering. In: *Proceedings of the 21st International Conference on Machine Learning*, pp 9–16
- Ben-Hur A, Noble W (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics* 21 Suppl 1:38–46
- Bhatia K, Jain H, Kar P, Varma M, Jain P (2015) Sparse local embeddings for extreme multi-label classification. In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp 730–738
- Bi W, Kwok J (2012) Mandatory leaf node prediction in hierarchical multilabel classification. In: *Advances in Neural Information Processing Systems 25*, pp 153–161
- Bi W, Kwok JT (2013) Efficient multi-label classification with many labels. In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp 405–413
- Bielza C, Li G, Larrañaga P (2011) Multi-dimensional classification with bayesian networks. *Int J Approx Reasoning* 52(6):705–727
- Bonilla EV, Agakov F, Williams C (2007) Kernel multi-task learning using task-specific features. In: *The 11th International Conference on Artificial Intelligence and Statistics AISTATS'07*, pp 43–50
- Breiman L, Friedman J (1997) Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society B* 69:3–54
- Candes E, Recht B (2008) Exact low-rank matrix completion via convex optimization. *Foundations of Computational Mathematics* 9:717–772
- Caponnetto A, Micchelli CA, Pontil M, Ying Y (2008) Universal multi-task kernels. *Journal of Machine Learning Research* 9:1615–1646
- Caruana R (1997) Multitask learning: A knowledge-based source of inductive bias. *Machine Learning* 28:41–75
- Chen J, Tang L, Liu J, Ye J (2009) A convex formulation for learning shared structures from multiple tasks. In: *Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA, ICML '09*, pp 137–144
- Cheng W, Hüllermeier E (2009) Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning* 76(2-3):211–225
- Cissé M, Usunier N, Artières T, Gallinari P (2013) Robust bloom filters for large multilabel classification tasks. In: *Advances in Neural Information Processing Systems 26, Lake Tahoe, Nevada, United States.*, pp 1851–1859
- Dembczyński K, Waegeman W, Cheng W, Hüllermeier E (2012) On label dependence and loss minimization in multi-label classification. *Machine Learning* 88:5–45
- Dembczyński K, Kotłowski W, Gawel P, Szarecki A, Jaszkievicz A (2013) Matrix factorization for travel time estimation in large traffic networks. In: *Artificial Intelligence and Soft Computing – 12th International Conference (ICAISC 2013)*, Springer, Lecture Notes in Computer Science, vol 7895, pp 500–510

- Dembczynski K, Kotlowski W, Waegeman W, Busa-Fekete R, Hüllermeier E (2016) Consistency of probabilistic classifier trees. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II, pp 511–526
- Deng J, Ding N, Jia Y, Frome A, Murphy K, Bengio S, Li Y, Neven H, Adam H (2014) Large-scale object classification using label relation graphs. In: European Conference on Computer Vision, Springer, Lecture Notes in Computer Science, vol 8689, pp 48–64
- Dinuzzo F (2013) Learning output kernels for multi-task problems. *Neurocomput* 118:119–126
- Dinuzzo F, Ong CS, Gehler P, Pillonetto G (2011) Learning output kernels with block coordinate descent. In: Proceedings of the International Conference on Machine Learning
- Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pp 647–655
- Evgeniou T (2005) Learning multiple tasks with kernel methods. *Journal of Machine Learning Research* 6:615–637
- Evgeniou T, Pontil M (2004) Regularized multi-task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 109–117
- Fang Y, Si L (2011) Matrix co-factorization for recommendation with rich side information and implicit feedback. In: The 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, ACM, pp 65–69
- Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T (2013) Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, pp 2121–2129
- Fu Y, Hospedales T, Xiang T, Gong S (2013) Learning multimodal latent attributes. *IEEE transactions on pattern analysis and machine intelligence* 36(2):303–316
- Gaujoux R, Seoighe C (2010) A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* 11:367
- Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp 580–587
- Godbole S, Sarawagi S (2004) Discriminative methods for multi-labeled classification. In: PAKDD 2004, pp 22–30
- Gönen M (2012) Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 28(18):2304–10
- Gong P, Ye J, Zhang C (2012) Robust multi-task feature learning. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '12, pp 895–903
- Gong Y, Wang L, Guo R, Lazebnik S (2014) Multi-scale orderless pooling of deep convolutional activation features. In: European Conference on Computer Vision, Springer, Lecture Notes in Computer Science, vol 8695, pp 392–407
- Gopal S, Yang Y (2013) Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp 257–265
- Gopal S, Yang Y, Bai B, Niculescu-Mizil A (2012) Bayesian models for large-scale hierarchical classification. In: Proceedings of the 25th International Conference on Neural

- Information Processing Systems, USA, NIPS'12, pp 2411–2419
- Gu Q, Li Z, Han J (2011) Correlated multi-label feature selection. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '11, pp 1087–1096
- Guo Y, Gu S (2011) Multi-label classification using conditional dependency networks. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, AAAI Press, IJCAI'11, pp 1300–1305
- Hariharan B, Zelnik-Manor L, Vishwanathan S, Varma M (2010) Large scale max-margin multi-label classification with priors. In: International Conference on Machine Learning, Omnipress
- Hastie T, Tibshirani R, Friedman JH (2007) Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer
- Hayashi K, Takenouchi T, Tomioka R, Kashima H (2012) Self-measuring similarity for multi-task gaussian process. In: Guyon I, Dror G, Lemaire V, Taylor GW, Silver DL (eds) ICML Workshop on Unsupervised and Transfer Learning, JMLR Proceedings, vol 27, pp 145–154
- Hsu D, Kakade S, Langford J, Zhang T (2009) Multi-label prediction via compressed sensing. In: NIPS 22, pp 772–780
- Hüllermeier E, Fürnkranz J, Cheng W, Brinker K (2008) Label ranking by learning pairwise preferences. *Artificial Intelligence* 172(16–17):1897–1916
- Izenman A (1975) Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* 5:248–262
- Jacob L, Vert J (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* 24(19):2149–2156
- Jacob L, Bach F, Vert JP (2008) Clustered multi-task learning: A convex formulation. *Advances in Neural Information Processing Systems*
- Jain P, Netrapalli P, Sanghavi S (2013) Low-rank matrix completion using alternating minimization. In: Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, ACM, New York, NY, USA, pp 665–674
- Jalali A, Sanghavi S, Ravikumar P, Ruan C (2010) A dirty model for multi-task learning. In: *Neural Information Processing Systems*, pp 964–972
- James W, Stein C (1961) Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability theory* p 361379
- Jawanpuria P, Lapin M, Hein M, Schiele B (2015) Efficient output kernel learning for multiple tasks. In: *Advances in Neural Information Processing Systems* 28, pp 1189–1197
- Kashima H, Kato T, Yamanishi Y, Sugiyama M, Tsuda K (2009) Link propagation: A fast semi-supervised learning algorithm for link prediction. In: *SIAM International Conference on Data Mining (SDM'09)*, SIAM, pp 1099–1110
- Kong X, Yu PS (2012) gmlc: a multi-label feature selection framework for graph classification. *Knowledge and Information Systems* 31(2):281–305
- Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: *Conference on Computer Vision and Pattern Recognition*
- Larochelle H, Erhan D, Bengio Y (2008) Zero-data learning of new tasks. In: *23rd national conference on Artificial intelligence (AAAI'08)*, AAAI Press, pp 646–651
- Lawrence N, Urtasun R (2009) Non-linear matrix factorization with Gaussian processes. In: *Proceedings of the 26th Annual International Conference on Machine Learning*
- Lee G, Yang E, Hwang SJ (2016) Asymmetric multi-task learning based on task relatedness and confidence. In: *Proceedings of the 33rd International Conference on Machine*

- Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, pp 230–238
- Liu H, Sun J, Guan J, Zheng J, Zhou S (2015) Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 31(12):i221–i229
- Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp 3337–3344
- Loza Mencía E, Janssen F (2016) Learning rules for multi-label classification: a stacking and a separate-and-conquer approach. *Machine Learning* 105(1):77–126
- Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research* 11:2287–2322
- Menon A, Elkan C (2010) A log-linear model with latent features for dyadic prediction. In: The 10th IEEE International Conference on Data Mining (ICDM), pp 364–373
- Menon A, Elkan C (2011) Link prediction via matrix factorization. *Machine Learning and Knowledge Discovery in Databases* 6912:437–452
- Van der Merwe A, Zidek J (1980) Multivariate regression analysis and canonical variates. *Canadian Journal of Statistics* 8:27–39
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. CoRR abs/1301.3781, URL <http://arxiv.org/abs/1301.3781>
- Nam J, Menca EL, Kim HJ, Fürnkranz J (2015) Predicting unseen labels using label hierarchies in large-scale multi-label learning. In: European Conference on Machine Learning, Springer, Lecture Notes in Computer Science, vol 9284, pp 102–118
- Nam J, Loza Mencía E, Fürnkranz J (2016) All-in text: Learning document, label, and word representations jointly. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., pp 1948–1954
- Obozinski G, Taskar B, Jordan MI (2010) Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing* 20(2):231–252
- Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp 1717–1724
- Oyama S, Manning C (2004) Using feature conjunctions across examples for learning pairwise classifiers. In: European conference on Machine learning and Knowledge Discovery in Databases, Springer, Lecture Notes in Computer Science, vol 3201, pp 322–333
- Pahikkala T, Waegeman W, Tsivtsivadze E, Salakoski T, De Baets B (2010) Learning intransitive reciprocal relations with kernel methods. *European Journal of Operational Research* 206(3):676–685
- Pahikkala T, Airola A, Stock M, Baets BD, Waegeman W (2013) Efficient regularized least-squares algorithms for conditional ranking on relational data. *Machine Learning* 93(2–3):321–356
- Pahikkala T, Stock M, Airola A, Aittokallio T, De Baets B, Waegeman W (2014) A two-step learning approach for solving full and almost full cold start problems in dyadic prediction. *Lecture Notes in Computer Science* 8725:517–532
- Palatucci M, Hinton G, Pomerleau D, Mitchell TM (2009) Zero-shot learning with semantic output codes. In: Advances in Neural Information Processing Systems, pp 1410–1418
- Papagiannopoulou C, Tsoumakas G, Tsamardinos I (2015) Discovering and exploiting deterministic label relationships in multi-label learning. In: Cao L, Zhang C, Joachims T, Webb GI, Margineantu DD, Williams G (eds) KDD, ACM, pp 915–924

- Park SH, Fürnkranz J (2008) Multi-label classification with label constraints. In: ECML PKDD 2008 Workshop on Preference Learning (PL-08, Antwerp, Belgium)
- Park ST, Chu W (2009) Pairwise preference regression for cold-start recommendation. In: The Third ACM Conference on Recommender Systems, ACM, pp 21–28
- Park Y, Marcotte EM (2012) Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods* 9(12):1134–1136
- Pelossof R, Singh I, Yang JL, Weirauch MT, Hughes TR, Leslie CS (2015) Affinity regression predicts the recognition code of nucleic acid-binding proteins. *Nature Biotechnology* 33(12):1242–1249
- Prabhu Y, Varma M (2014) Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp 263–272
- Rai P, Daumé III H (2009) Multi-label prediction via sparse infinite CCA. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS)
- Rangwala H, Naik A (2017) Large scale hierarchical classification: foundations, algorithms and applications. KDD Tutorial, Halifax, Canada
- Raymond R, Kashima H (2010) Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. In: Balcázar JL, Bonchi F, Gionis A, Sebag M (eds) European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Lecture Notes in Computer Science, vol 6323, pp 131–147
- Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: An astounding baseline for recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp 512–519
- Read J (2013) Multi-Dimensional Classification with Super-Classes. *IEEE Transactions on Knowledge and Data Engineering* 99:1
- Rohrbach M, Stark M, Schiele B (2011) Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 1641–1648
- Romera-Paredes B, Torr P (2015) An embarrassingly simple approach to zero-shot learning. In: Proceedings of the 32nd International Conference on Machine Learning, vol 37, pp 2152–2161
- Rousu J, Saunders C, Szedmak S, Shawe-Taylor J (2006) Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research* 7:1601–1626
- Schäfer D, Hüllermeier E (2015) Dyad ranking using a bilinear Plackett-Luce model. In: Proceedings ECML/PKDD–2015, European Conference on Machine Learning and Knowledge Discovery in Databases, Porto, Portugal
- Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, Lecun Y (2014) Overfeat: Integrated recognition, localization and detection using convolutional networks
- Shan H, Banerjee A (2010) Generalized probabilistic matrix factorizations for collaborative filtering. In: Webb GI, Liu B, Zhang C, Gunopulos D, Wu X (eds) The 10th IEEE International Conference on Data Mining (ICDM), IEEE Computer Society, pp 1025–1030
- Shawe-Taylor J, Cristianini N (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press
- Silla CN, Freitas Aa (2010) A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22(1-2):31–72
- Socher R, Ganjoo M, Sridhar H, Bastani O, Manning CD, Ng AY (2013) Zero-shot learning through cross-modal transfer. CoRR abs/1301.3666
- Spolař N, Monard MC, Tsoumakas G, Lee HD (2016) A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomputing*

- 180(C):3–15
- Stock M, Fober T, Hüllermeier E, Glinca S, Klebe G, Pahikkala T, Airola A, De Baets B, Waegeman W (2014) Identification of functionally related enzymes by learning-to-rank methods. *IEEE Transactions on Computational Biology and Bioinformatics* 11(6):1157–1169
- Stock M, Pahikkala T, Airola A, Baets BD, Waegeman W (2016) Efficient pairwise learning using kernel ridge regression: an exact two-step method. *Arxiv abs/1606.04275*
- Tai F, Lin HT (2010) Multi-label classification with principle label space transformation. In: *Second International Workshop on Learning from Multi-Label Data (MLD 2010)*, in conjunction with ICML/COLT 2010
- Tai F, Lin HT (2012) Multilabel classification with principal label space transformation. *Neural Computation* 24(9):2508–2542
- Takács G, Pilászy I, Németh B, Tikk D (2008) Matrix factorization and neighbor based algorithms for the netflix prize problem. In: *Proceedings of the 2008 ACM conference on Recommender systems*, ACM Press, New York, New York, USA, pp 267–274
- Tsoumakas G, Katakis I (2007) Multi label classification: An overview. *Int J Data Warehousing and Mining* 3(3):1–13
- Van Loan CF (2000) The ubiquitous kronecker product. *Journal of Computational and Applied Mathematics* 123(1–2):85–100
- Van Peer G, Paepé AD, Stock M, Anckaert J, Volders PJ, Vandesompele J, Baets BD, Waegeman W (2017) mistar: mirna target prediction through modeling quantitative and qualitative mirna binding site information in a stacked model structure. *Nucleic Acids Research* 45:e51
- Vens C, Struyf J, Schietgat L, Dzeroski S, Blockeel H (2008) Decision trees for hierarchical multi-label classification. *Machine Learning* 73(2):185–214
- Vert JP, Qiu J, Noble WS (2007) A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics* 8(S-10):1–10
- Volkovs M, Zemel RS (2012) Collaborative ranking with 17 parameters. In: *Advances in Neural Information Processing Systems*, pp 2303–2311
- Waegeman W, Pahikkala T, Airola A, Salakoski T, Stock M, De Baets B (2012) A kernel-based framework for learning graded relations from data. *IEEE Transactions on Fuzzy Systems* 20(6):1090–1101
- Waegeman W, Dembczynski K, Jachnik A, Cheng W, Hüllermeier E (2014) On the bayes-optimality of f -measure maximizers. *Journal of Machine Learning Research* pp 3333–3388
- Wang F, Wang X, Li T (2009) Semi-supervised multi-task learning with task regularizations. In: *IEEE International Conference on Data Mining*, pp 562–568
- Wei Y, Xia W, Lin M, Huang J, Ni B, Dong J, Zhao Y, Yan S (2016) Hcp: A flexible cnn framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(9):1901–1907
- Weston J, Chapelle O, Elisseeff A, Schölkopf B, Vapnik V (2002) Kernel dependency estimation. In: *Advances in Neural Information Processing Systems*, pp 873–880
- Weston J, Makadia A, Yee H (2013) Label partitioning for sublinear ranking. In: *Proceedings of the 30th International Conference on Machine Learning*
- Wicker J, Tyukin A, Kramer S (2016) A nonlinear label compression and transformation method for multi-label classification using autoencoders. In: *Advances in knowledge discovery and data mining : 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand*

- Wolpert DH (1992) Original contribution: Stacked generalization. *Neural Networks* 5(2):241–259
- Xian Y, Akata Z, Sharma G, Nguyen QN, Hein M, Schiele B (2016) Latent embeddings for zero-shot classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp 69–77
- Xian Y, Schiele B, Akata Z (2017) Zero-shot learning - the good, the bad and the ugly. *CoRR*
- Xue Y, Liao X, Carin L, Krishnapuram B (2007) Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research*, 8:3563
- Yen IE, Huang X, Ravikumar P, Zhong K, Dhillon IS (2016) Pd-sparse : A primal and dual sparse approach to extreme multiclass and multilabel classification. In: *Proceedings of the 33rd International Conference on Machine Learning*, New York City, NY, USA, pp 3069–3077
- Yu S, Yu K, Tresp V, Kriegel HP (2006) Multi-output regularized feature projection. *IEEE Transactions on Knowledge and Data Engineering* 18(12):1600–1613
- Zhang D, Shen D (2012) Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer’s disease. *NeuroImage* 59(2):895–907
- Zhang Y, Schneider J (2011) Multi-label Output Codes using Canonical Correlation Analysis. In: *Uncertainty in Artificial Intelligence*
- Zhang Y, Yeung D (2010) A convex formulation for learning task relationships in multi-task learning. *UAI abs/1203.3536*
- Zhou J, Chen J, Ye J (2011a) Clustered multi-task learning via alternating structure optimization, chap *Advances in Neural Information Processing Systems*
- Zhou J, Yuan L, Liu J, Ye J (2011b) A multi-task learning formulation for predicting disease progression. In: Apt C, Ghosh J, Smyth P (eds) *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp 814–822
- Zhou J, Liu J, Narayan VA, Ye J (2012a) Modeling disease progression via fused sparse group lasso. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, KDD ’12, pp 1095–1103
- Zhou T, Shan H, Banerjee A, Sapiro G (2012b) Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In: *12th SIAM International Conference on Data Mining*, SIAM, pp 403–414
- Zhou Z, Zhang M (2007) Multi-instance multilabel learning with application to scene classification. In: *Advances in Neural Information Processing Systems* 19