A snapshot on nonstandard supervised learning problems

III.

Source

Charte, D., Charte, F., García, S., & Herrera, F. (2019). A snapshot on nonstandard supervised learning problems: taxonomy, relationships, problem transformations and algorithm adaptations. *Progress in Artificial Intelligence, 8(1)*, 1-14.

Abstract

Machine learning is a field which studies how machines can alter and adapt their behavior, improving their actions according to the information they are given. This field is subdivided into multiple areas, among which the best known are supervised learning (e.g. classification and regression) and unsupervised learning (e.g. clustering and association rules).

Within supervised learning, most studies and research are focused on well known standard tasks, such as binary classification, multiclass classification and regression with one dependent variable. However, there are many other less known problems. These are what we generically call nonstandard supervised learning problems. The literature about them is much more sparse, and each study is directed to a specific task. Therefore, the definitions, relations and applications of this kind of learners are hard to find.

The goal of this paper is to provide the reader with a broad view on the distinct variations of nonstandard supervised problems. A comprehensive taxonomy summarizing their traits is proposed. A review of the common approaches followed to accomplish them and their main applications is provided as well.

Keywords

Machine learning - Supervised learning - Nonstandard learning

III.1. Introduction

According to Mitchell [1], a machine is said to learn from experience E related to a class of tasks T and performance metric P, when its performance at tasks in T improves according to P after experience E.

Supervised learning is one of the fundamental areas of machine learning [2]. From object detection to ecological modeling to emotion recognition, it covers all kinds of applications. It essentially consists in learning a function by training with a set of input-output pairs. The training stage can be seen as E in the previous definition, and the specific task T may vary, but usually involves predicting an appropriate output given a new input.

Traditionally, supervised learning problems have been spread into two categories: classification and regression [3, 4]. In the first, information is divided into discrete categories, while the latter involves patterns associated to a value in a continuous spectrum.

These problems can be processed by learning from a training dataset, which is composed of instances. Typically, these instances or samples take the form x, y where x is a vector of values in the space of input variables and y is a value in the target variable. Each problem can be described by the type of its instances: inputs will usually belong to a subset of $\mathbb{R}^n$, and outputs will take values in a specific one-dimensional set, finite or continuous. Once trained, the obtained model can be used to predict the target variable on unseen instances.

Standard classification problems are those where labels are either binary or multiclass [5, 6]. In the binary case, an instance can only be associated with one of two values: positive or negative, which is equivalent to 0 or 1. For example, email messages may be classified into spam or legit, and tumours can be categorized as either benign or malign. Multiclass problems, on the other hand, involve any finite number of classes. That is, any given instance will belong to one of possibly many categories, which is equivalent to it being assigned a natural number below a convenient threshold. As an example, a photograph of a plant or a sound recording from an animal could correspond to one of a variety of species.

A standard regression problem [7, 8] consists in finding a function which is able to predict, for a given example, a real value among a continuous range, usually an interval or the set of real numbers $\mathbb{R}$. For example, the height of a person may be estimated out of several characteristics such as age or country of origin.

Even though these standard problems are applicable in a multitude of cases, there are situations whose correct modeling requires modifications of their structure. For example, a newspaper article can be categorized according to its contents, but it could be desirable to assign several categories simultaneously. Similarly, a social media post could be described by not one but two input vectors, an image and a piece of text. These special circumstances cannot be covered by the traditional one-vector input and one-dimensional output schema. As a consequence, since performance metrics which

measure improvements in standard tasks assume the common structure, they lose applicability or sense in these cases. Thus, not only new techniques are needed to tackle the problems, but also new ways of measuring and comparing their success.

This work studies variations on classic supervised problems where the traditional structure is not obeyed, which we call nonstandard variations. These emerge when the structure of the classical components of the problems does not suffice to describe complex situations, such as multiplicity of inputs or outputs, or order restrictions. As a consequence, this manuscript does not cover other singular supervised problems, such as high dimensionality of the feature space [9] or unbalanced training sets [10, 11], nor time-dependent problems, such as data streams [12, 13] or time series [14].

The rest of the paper is structured as follows. Section III.2 formally defines and describes each nonstandard variation. This is followed by Section III.3 establishing relations among the introduced problems and proposing a taxonomy of them. Section III.4 describes the most common techniques used to solve them. After that, Section III.5 enumerates popular applications of each problem. Section III.6 covers other variations further from the ones previously detailed. Lastly, Section III.7 draws some conclusions.

III.2. Definitions of nonstandard variations

The problems introduced in this section are generalizations over the traditional versions of classification and regression. The focus is on fully supervised problems, where inputs are always paired with outputs during training. An alternative taxonomy based on different supervision models is introduced in [15].

Notation

In this work we will establish a notation which intends to be as simple to understand as possible, while being able to encompass every nonstandard variation. First, any supervised learning problem consists in finding a function which will classify, rank or perform regression. It will be noted as

$$f : X \to Y \tag{III.1}$$

where X is an input set, or domain, and Y is an output set, or codomain. It will be assumed that a training dataset S is provided,

including a finite number of input-output pairs:

$$x, y \in S \subset X \times Y \ . \tag{III.2}$$

This way, a learning algorithm will be able to generate the desired function f. An additional notation will be the set of labels $\mathcal{L}$ where convenient.

For example, in standard binary classification $X \subset \mathbb{R}^n$ and $Y = \mathcal{L} = \{0, 1\}$. Similarly, standard regression problems can be defined with the same kind of X set and $Y \subset \mathbb{R}$. Thus, we can define very distinct supervised problems by particularizing sets X or Y in different ways.

Other usual notations are based in probability theory, thus involving random variables and probability distributions [16, 17]. In that case, X and Y would be the sample spaces of the input and output variables $\mathbf{X}$ and $\mathbf{Y}$, respectively. Predictors would usually attempt to infer a discriminant model $P\mathbf{Y}|\mathbf{X}$ from the training dataset.

Multi-instance

The multi-instance (MI) framework [18] assumes a single feature space for all instances, but each training pattern may consist of more than one instance. In this case, a training pattern is composed of a finite multiset or *bag* of instances and a label. Formally, assuming instances are drawn from a set $A \subset \mathbb{R}^n$, the domain can be described as follows:

$$X = \{b \subset A \mid b \text{ finite}\} \ . \tag{III.3}$$

In this case, the learning algorithm will not know labels associated to each instance but to a bag of them. In addition to this, not all instances may share the same relevance or are equally related to the label.

Some MI problems assume that hidden labels are present for each instance in a bag: for example, a training set of drug tests where, for each test, several drug types are analyzed. Additionally, a typical MI assumption in the binary scenario states that a bag is positive when at least one of its instances is positive, and it is negative otherwise [19].

Other MI problems differ in that a per-instance labeling may not be possible or may not make sense: for example, if each bag represents an image and instances are image segments, class *beach* can only apply to bags with water and sand segments, but it cannot apply to an individual instance.

Multi-view

A learning problem is considered to be multi-view (MV) [20] when inputs are composed of several components of very different nature.

For example, if a learning pattern consists of an image as well as a piece of text representing the same instance, they can be seen as two *views* on it. In that case, images and texts would belong to distinct feature spaces A and B respectively, an input pattern being $a, b \in A \times B$. More generally, we can describe the input space as:

$$X = \prod_{i=1}^{t} A_i \text{ , where } A_i \subset \mathbb{R}^{n_i}, \tag{III.4}$$

where t is the number of views offered by the problem and n_i is the dimension of the feature space of the i-th view.

Multi-label

The multi-label (ML) learning field [21, 22] studies problems related to simultaneously assigning multiple labels to a single instance. That is, if $\mathcal{L} = \{l_1, \ldots, l_p\}$ the codomain consists of all possible selections of these p labels, also known as *labelsets*:

$$Y = 2^{\mathcal{L}} \cong \{0, 1\}^p \ . \tag{III.5}$$

As shown by this formulation, it is equivalent to think of a selection of labels as a subset of $\mathcal{L}$ and as a binary vector. For example, the labelset composed of the first and third labels can be represented either by $\{l_1, l_3\}$ or $1, 0, 1, 0, \ldots, 0$.

The difference that arises when comparing ML problems to binary or multiclass ones is that labels may interact with each other. For example, a news piece classified in *economy* is more likely to be labeled *politics* than *sports*. Similarly, a photograph labeled *ocean* is less likely to have the *mountains* label rather than *beach*. Methods may take advantage of label co-ocurrence [23] in order to reduce the search space when predicting a labelset.

A constrained version of ML classification is hierarchical ML classification [24], where labels are organized in a class hierarchy, usually a tree or a direct acyclic graph. A predicted labelset for a given instance is only consistent if parents of all labels in the labelset are also predicted.

Multi-dimensional

Multi-dimensional (MD) learning [25] is a generalized classification problem where categorization is performed simultaneously along several dimensions. Each instance can belong to one of many classes in each dimension, thus the output space can be formally described as:

$$Y = \mathcal{L}_1 \times \mathcal{L}_2 \times \cdots \times \mathcal{L}_p, \tag{III.6}$$

where $\mathcal{L}_i$ is the label space for the i-th dimension.

As with ML learning, label dimensions may be related in some way and treating them independently would only be a naive solution to the problem.

Label distribution learning

In label distribution learning (LDL) problems [26], otherwise known as probabilistic class label problems [27], any instance can be described in different degrees by each label. This can be modeled as a discrete distribution over the labels, where the probability of a label given a specific instance is called its *degree of description*. Analitically, the objective is, for each instance, to predict a real-valued vector which sums exactly 1:

$$Y = \left\{ y \in \left[0,1\right]^p : \sum_{i=1}^{p} y_i = 1 \right\}. \tag{III.7}$$

In this case, we would say that the i-th label in $\mathcal{L}$ describes an instance x, y with degree y_i.

Label ranking

In a label ranking (LR) problem [28, 29] the objective is not to find a function able to choose one or several labels from the label space. Instead, it must evaluate their relevance for each unseen instance. The most general version of the problem involves a training set where Y is the set of all partial orders of $\mathcal{L}$, and the obtained function also maps individual instances to partial orders. This way, for each test instance the function will output a sequence of preferences where some labels will be seen as more relevant than others.

However, the typical situation in label ranking problems is that the orders are total, which means any two labels can always be compared. This is called a *ranking* and does not exclude the

possibility of ties. When ties are not allowed it is said to be a *sorting* or *permutation*, and can be formulated as follows:

$$Y = \{\sigma : \{1, \ldots, p\} \to \mathcal{L} \mid \sigma \text{ is bijective}\} \ , \tag{III.8}$$

where p is the amount of labels. Y can also be seen as the set of all permutations of the labels in $\mathcal{L}$, usually known as the symmetric group of order p, and noted as S_p.

Multi-target regression

A regression problem where the output space has more than just one dimension is usually called multi-target regression (MTR) and is also known as multi-output, multi-variate or multi-response [30]. In this case, a formal description is simply that the codomain is a continuous multi-dimensional real set:

$$Y = \prod_{i=1}^{p} Y_i \text{ , where } Y_i \subset \mathbb{R} \ \forall i \tag{III.9}$$

and p is the number of target variables.

As with other multiple target extensions, the key difference with single-target regression in this case is the possible interactions among output variables.

Ordinal regression

A problem where the target space is discrete but ordered is called ordinal regression (OR) or, alternatively, ordinal classification [31]. It can be located midway between classification and regression. More specifically, it consists in labeling instances with a finite number of choices where these are ordered

$$Y = \{1, 2, \ldots, c\}, \ 1 < 2 < \cdots < c \ . \tag{III.10}$$

In OR, the training phase consists in learning from a set of feature vectors which have a specific label associated to them, and testing can be performed over individual instances. This means that, although labels are ordered, the main objective is not to rank or sort instances as in learning to rank [32], but to simply classify them. The labels themselves do not provide any metric information either, they only carry qualitative information about the order among themselves.

Monotonicity constraints

Order relations can exist not only in the label space but in the feature space as well. Partial orders among real-valued feature vectors are always possible, and there may be cases where the order among instances is determined by just one or a few of their attributes.

When inputs as well as outputs are at least partially ordered, it is common to look for predictions which respect their order relations. In that case, the objective is to obtain a classifier or regression function which enforces the following constraint:

$$x_1 < x_2 \Rightarrow fx_1 < fx_2 \; \forall x_1, x_2 \in X \; . \tag{III.11}$$

When Y is discrete the problem is usually called *monotone classification* (MC), monotonic classification or ordinal classification with monotonicity constraints [33]. If, on the contrary, Y is continuous, it is known as *isotonic regression* (IR) [34].

Absence or partiality of information

Some problems do not directly alter the structure of X and Y from the standard supervised problem. Instead, they restrict which data can belong to a training set, or remove labelings from training examples. In this case, training information is presented partially or with some exclusions.

According to which kind of information is missing from the training set, a learning task can usually be categorized as semi-supervised [35], one-class learning [36], PU-learning [37], zero-shot learning [38] or one-shot learning [39]. These are described further in Section III.6.

Variation combinations

Some of the components described above can be combined to compose a more complex problem overall. Usually, one of these combinations will take components from different variation types, for example, simultaneous multiplicity of inputs and outputs.

More specifically, there exist several studies involving MI ML scenarios [40, 41]. In this case, examples from the input space are composed of several feature vectors and are associated to various labels. As a consequence, this model can represent many complicated problems where inputs and outputs have more structure than usual.

Other more uncommon situations are MV MI ML problems [42], where patterns have several instances which may or may not belong to the same space, a multi-output version of OR named graded ML classification [43] and more complex input structures such as multi-layer MI MV [44], where a hierarchy of instances is present in each example.

III.3. Taxonomy

A first categorization of the variations analyzed in this work can be made according to how they differ from the standard problem. There can be multiplicity in the input space or the output space, order constraints may exist, or only partial information may be given in some cases. Fig. III.1 shows ways in which the traditional problems can be generalized.

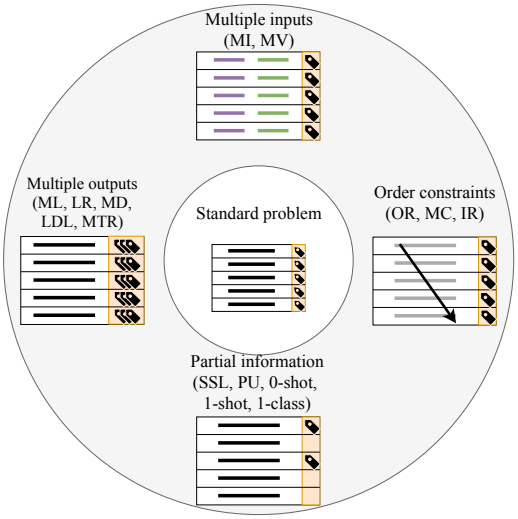

Figure III.1.: Extensions of the standard supervised problem: multiple inputs or outputs, presence of orders and rankings, and partial information.

Problems introducing multiple inputs are MI and MV, whereas multiple outputs can be found on ML, MD, LR, LDL and MTR. Problems where orders are present are OR, MC and IR. Likewise, tasks with only partial information are, among others, semi-supervised learning (SSL), positive-unlabeled (PU) learning, one-shot classification and zero-shot classification.

Finally, a generalized problem can be built out of combining several of these components: for example, a multiple-input multiple-output problem where the inputs and outputs can belong to structures like the ones defined above.

The rest of this section studies variations on the structure of the input space and output space, establishes relations among problems, and describes how they can be particularized or generalized to one another.

Input structure

In a standard supervised problem, the input space consists of single feature vectors and does not impose a specific order.

Problems where learning patterns are composed of multiple instances can usually be categorized into either MI, if the inputs share the same structure, or MV, otherwise. Their combination can also be considered as well, e.g. a problem where an example is composed of one or more photographs and one or more pieces of text. This would be a case of a MV MI problem.

There are also problems where there exists a partial or total order among instances, which is coupled with an order constraint in relation to the outputs. These are MC and IR.

Fig. III.2 summarizes these structural traits in a hierarchy and indicates problems where these traits are present.

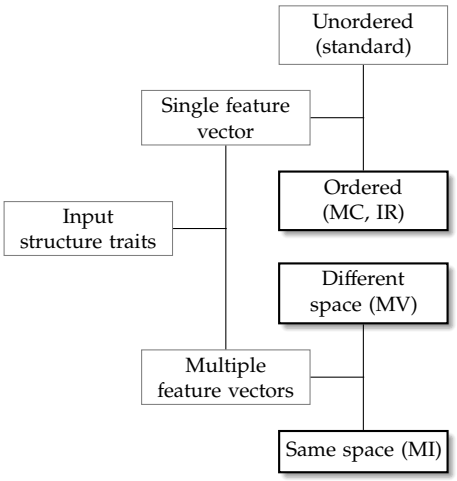

Figure III.2.: Traits that can be found on the input structure of supervised problems.

Output structure

The diversity in output variations is higher than that of the input ones. A first sorting criterion is whether the codomain is discrete or continuous. This way, problems are either classification or regression ones.

Further subdivision of problems allows to separate these traits according to whether outputs remain scalars or become vectors. In the first case we consider order in the discrete scenario a nonstandard variation, which is present in OR and MC. In the second case, classification problems are spread into ML, LR and MD, and regression ones into LDL and MTR.

Fig. III.3 organizes these traits in a hierarchy based on the previous criteria. Each leaf of the tree also includes problems where each one is present.

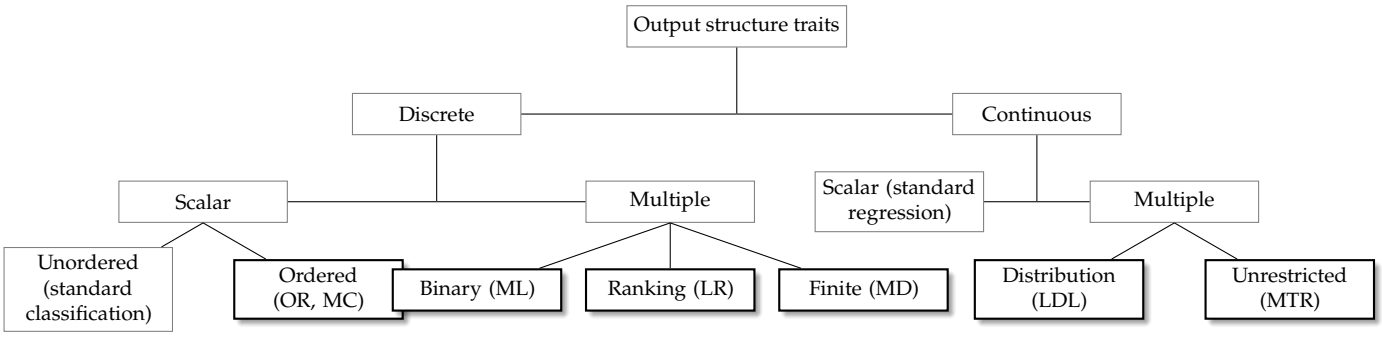

Figure III.3.: Traits that can be found on the output structure of supervised problems.

The variations in the structure of target spaces in supervised problems can be seen as generalizations of the standard problems. Furthermore, some of them are also more general than others. For example, ML problems can be seen as LR ones where, for a given instance, labels over a threshold are active and those below are not. Thus, LR is a generalization of the ML scenario. More relations of this kind are displayed in Fig. III.4.

As shown in the graph, an inclusion of more target variables of the same type transforms a binary problem into ML, a multiclass problem into MD and a single-target regression one into MTR. Similarly, inclusion of more values into each variable allows to generalize binary problems to multiclass, and ordinal to single-target regression, as well as ML ones to MD and these to MTR. LDL can be seen as a generalization of ML where real numbers between 0 and 1 are also allowed as values for a label. LR is a generalization of ML by the argument discussed before.

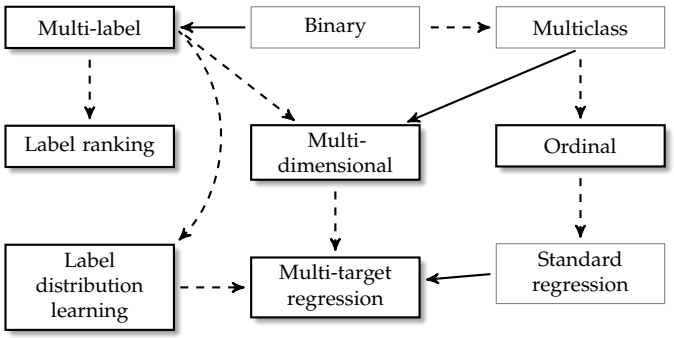

Figure III.4.: Relations among supervised problems according to output structure. Arrows follow natural generalizations from one problem to another. Continuous arrows denote generalizations based on adding more variables of the same type. Dashed arrows indicate generalizations based on modifying existing target variables.

Summary

In this section input and output variations of standard supervised problems have been categorized and related. Table III.1 allows to identify specific problems according to which input and output traits are present.

Table III.1.: Identification of problems according to their input traits (vertical axis) and output traits (horizontal axis).

Inputs \ Outputs	Unordered outputs		Ordered outputs			
	Scalar	Multiple	Scalar		Multiple	
			Discrete	Continuous	Discrete	Continuous
Unordered inputs	standard classification [3]	ML/MD classification [21, 25]	OR [31]	standard regression [8]	Graded ML [43]	MTR [30]
Ordered inputs	-	-	MC [33]	IR [34]	-	-
Multiple instances	MI classification [18]	MIML/MIMD classification [40]	-	MI regression [18]	-	-
Multiple views	MV classification [20]	MVML/MVMD classification [42]	-	MV regression [20]	-	-

III.4. Common approaches to tackle nonstandard problems

When tackling a nonstandard problem, most techniques follow one of two main approaches: problem transformation or algorithm adaptation. The first one relies on appropriate transformations of the data which result in one or more simpler, standard problems. The latter implies an extension or development of previously existing algorithms, in order to adapt them to the complexities induced by the structure of the data.

In the following subsections several methods based on both approaches are enumerated for each analysed problem.

Problem transformation

Problem transformation methods assume that a solution can be achieved by extracting one or more simpler problems out of the original one. For example, a problem with multi-dimensional targets could be transformed into many problems with scalar outputs. Then, these problems could be solved independently by a classical algorithm. A solution for the original problem would be the concatenation of those extracted from the simpler ones.

Next, the most common transformation techniques are described for each nonstandard supervised learning task previously introduced.

– MI. The taxonomy proposed in [45] describes an Embedded Space paradigm, where each bag is transformed into a single feature vector representing the relevant information about the whole bag. This transformation brings the MI problem into a single-instance one. Most of these methods are vocabulary-based, which means that the embedding uses a set of concepts to classify each bag according to its instances, resulting in a single vector with one component per concept.

– MV. Some naive transformations consist in ignoring every view except one, or concatenating feature vectors from all views, thus training a single-view model in both cases [46]. A preprocessing based on Canonical Correlation Analysis [47] is able to project data from multiple views onto a lower-dimensional, single-view space.

– ML. Transformation methods for ML classification [48] are diverse: Binary Relevance trains separate binary classifiers for each label. Label Powerset reduces the problem to a multiclass one by treating each individual labelset as an independent class label, and Random k-Labelsets [49] extracts an ensemble of multiclass problems similarly. Classifier chains [50] trains subsequent binary classifiers accumulating previous predictions as inputs. ML problems can also be transformed to LR [51].

– MD. In some cases, independent classifiers can be trained for several dimensions [25, 52] but this method ignores possible correlations among dimensions. An alternative transformation, building a different label from each combination of classes, would produce a much larger label space and thus is not typically applied.

– LDL. A LDL problem can be reduced to multiclass classification by extracting as many single-label examples as labels for each one of the training instances [26]. These new examples are assigned a class corresponding to each label and weighted according to its degree of description. During the prediction process, the classifier must be able to output the score/confidence for each label, which can be used as its description degree.

– LR. A reduction of this problem to several binary problems can be achieved by learning pairwise preferences [28]. This transforms a c-label problem into $cc - 12$ binary problems describing a comparison among two labels. An alternative reduction by means of constraint classification [53] builds a single binary classification dataset by expanding each label preference into a new positive instance and a new negative instance. The feature space of the new binary problem has dimension nc, where n is the original dimension and c the number of labels, due to the constraints embedded in it by Kesler's construction [54].

– MTR. There are several ways to transform a MTR problem into several single-target regression ones. Some of them are inspired

by the ML field, such as a one-vs-all single-target reduction, multi-target stacking and regressor chains [55]. All of them train single-target regressors for several extracted problems, and then combine the obtained predictions. A different approach based on support vectors [56] extends the feature space which expresses the multi-output problem as a single-target one that can be solved using least squares support vector regression machines.

– OR. An ordinal problem with c classes can be transformed into $c-1$ binary classification problems by using each class from the second to the last one as a threshold for the positive class [57]. This decomposition can be called *ordered partitions* and is not the only possible one: others are *one-vs-next*, *one-vs-followers* and *one-vs-previous* [31]. Several 3-class problems can also be obtained by using, for the i-th problem, classes "l_i", "$< l_i$" and "$> l_i$".

– MC. The authors in [58] describe a procedure to tackle binary MC problems by means of IR. Multiclass MC cases can be reduced to several binary MC ones, which in turn are solved as IR problems.

Algorithm adaptation

Existing methods for classical problems can be extended in order to introduce the necessary complexities of nonstandard variations. As an example, nearest neighbor methods could be coupled with new distance metrics in order to be able to measure similarity among multiple inputs.

The rest of this section presents some algorithm adaptations which can be used to tackle nonstandard supervised tasks.

– MI. Methods that work on instance level are adaptations of algorithms from single-instance classification whose responses are then aggregated to build the bag-level classification [45]. They typically assume that one positive instance implies a positive bag. Adaptations of common algorithms have been proposed with support vector machines (SVM) [59] and neural networks [60], whereas some original methods in this area are Axis-Parallel Rectangles [61] and Diverse Density [62]. In the bag-space paradigm, methods treat bags as a whole and use specific distance metrics with distance as well as kernel-based classifiers, such as k-nearest neighbor (k-NN) [63] or SVM [64].

– MV. Supervised methods for MV are comparatively less developed than semi-supervised ones. Nonetheless, there is an extension of SVM [65] which simultaneously looks for two SVMs, one in each of the feature spaces of a two-view problem. There is an extension of Fisher discriminant analysis as well [66].

– ML. The most relevant algorithm adaptations [48] are based on standard classification algorithms with added support for choosing more than one class at a time: adaptations exist for k-NN [67], decision trees [68], SVMs [69], association rules [70] and ensembles [71].

– MD. Specific Bayesian networks have been proposed for the MD scenario [72, 73], as well as Maximum Entropy-based algorithms [25, 52].

– LDL. Proposals in [26] are adaptations of k-NN, with a special derivation of the label distribution of an unseen instance given its neighbors, and backpropagated neural networks, where the output layer indicates the label distribution of an instance. Other proposed methods are based on the optimization algorithms BFGS and Improved Iterative Scaling.

– LR. Boosting methods have been adapted to LR [74], as well as the SVM proposed in [69] for ML which can be naturally extended to LR [29]. An adaptation of online learning algorithms such as the perceptron has also been developed [75].

– MTR. First methods able to treat MTR problems were actually generalizations of statistical methods for single-target regression [76, 77]. Other common methods which have been extended to predict multiple regression variables are support vector regression [78, 79], kernel-based methods [80, 81], and regression trees [82] as well as random forests [83].

– OR. Neural networks can be used to tackle OR with slight changes in the loss function or the output layer [84, 85]. Similarly, extreme learning machines have also been applied to this problem [86, 87]. Common techniques such as k-NN or decision trees have been coupled with global constraints for OR [88], and extensions of other well known algorithms such as Gaussian processes [89] and AdaBoost [90] have been proposed as well.

– MC. Algorithm adaptations generally take a well known technique and add monotonicity constraints. For example, there exist in the literature adaptations of k-NN [91], decision trees [92], decision rules [93, 94] and artificial neural networks [95].

Table III.2 gathers all the methods described previously to tackle nonstandard supervised tasks.

Table III.2.: Summary table of presented methods according to their type of approach.

Task	Problem transformation	Algorithm adaptation
MI	Embedded-space [45]	SVM [59, 64] Neural networks [60] k-NN [63]
MV	Canonical correlation analysis [47]	SVM [65] Fisher discriminant analysis [66]
ML	Binary Relevance [48] Label Powerset [48] Classifier chains [50]	k-NN [67] Decision trees [68] SVM [69] Association rules [70] Ensembles [71]
MD	Independent classifiers [25, 52]	Bayesian networks [72, 73] Maximum Entropy [25, 52]
LDL	Multiclass reduction [26]	k-NN [26] Neural networks [26]
LR	Pairwise preferences [28] Constraint classification [53]	Boosting [74] SVM [29] Perceptron [75]
MTR	ML inspired: one-vs-all, stacking, regressor chains [55] Support vectors [56]	Generalizations [76, 77] Support vector regression [78, 79] Kernel-based [80, 81] Regression trees [82] Random forests [83]
OR	Ordered partitions [57] One-vs-next, One-vs-followers, One-vs-previous [31] 3-class problems [31]	Neural networks [84, 85] Extreme learning machines [86, 87] Decision trees [88] Gaussian processes [89] AdaBoost [90]
MC	Reduction to IR [58]	k-NN [91] Decision trees [92] Decision rules [93, 94] Neural networks [95]

III.5. Applications. Original real word scenarios

The problems studied in this work have their origins in real-world scenarios which are related below:

– MI. Problems modeled under MI learning are drug activity prediction [61], where each pattern describes a molecule and its different forms are represented by instances; image classification [45], and bankruptcy [96]. Most of the datasets used in experimentations, however, are usually synthetic.

– MV. Some situations where data is described in multiple views are multilingual text categorization [97], face detection with several poses [98], user localization in a WiFi network [99], advertisements described by their image and surrounding text [100] and image classification with several color-based views and texture-based views [101].

– ML. Problems which fall naturally under the ML definition are text classification under several categories simultaneously [102], image labeling [103], question tagging in forums where tags can co-exist [104], protein classification [105], data streams [106] and recommendation systems [107].

– MD. Applications of MD classification include classification of biomedical text [25], where predicted dimensions for a given document are its focus, evidence type, certainty level, polarity and trend; gene function identification [72]; tumor classification, and illness diagnosis in animals [73].

– LR. The field known as *preference learning* has been gaining interest [28], and LR is one of the problem that falls under this term. LR is also frequently applied in ML scenarios [108], where a threshold can be applied in order to transform an obtained ranking into a labelset.

– LDL. Data with relative importance of each label appears in applications such as analysis of gene expression levels in yeast [109], or emotion description from facial expressions [110], where a face can depict several emotions in different grades.

– MTR. Applications modeled as MTR problems are diverse, including modeling of vegetation condition in ecosystems assigning several scores which depend on the vegetation type [111], prediction of audio spectrums of wind tunnel tests [112], and estimation of several biophysical parameters from remote sensing images [113].

– OR. The most salient fields where OR can be found are text classification [114], where the predicted variable may be an opinion scale or a degree of satisfaction; image categorization [115]; medical research [116]; credit rating [117], and age estimation [118].

– MC. Monotonicity constraints are found in problems related to customer satisfaction analysis [119], in which overall appreciation of a product must increase along with the evaluation of its features; house pricing [92]; bankruptcy risk evaluation [120], and cancer prediction [121], among others.

III.6. Other nonstandard variations

This section covers variations of the standard supervised problem which are further from the central focus of this paper less related to those above.

Learning with partial information

In a standard supervised classification setting, it is assumed that every training example is labeled accordingly and that there exist examples for every class that may appear in the testing phase. When only a fraction of the training instances are labeled, the problem is considered semi-supervised [35], but generally there still exist labeled samples for each class.

In positive-unlabeled learning [37, 122], however, labeled examples provided within the training set are only positive. This means the learning algorithm only knows about the class of positive instances, and unlabeled ones can have either class.

A different scenario arises when the training set only consists of negative (or only positive) instances, and no unlabeled examples are provided. This is known as one-class classification [36], and data of this nature can be obtained from outlier detection applications, where positive examples are hardly recorded.

A problem which may be seen as a generalization of one-class classification is zero-shot learning [38], a situation where unseen classes are to be predicted in the testing stage. That is, the label space Y includes some values which are not present in any training pattern, but the classifier must be able to predict them. For example, if in a speech recognition problem Y is the set of all words in English, the training set is unlikely to have at least one instance for each word, thus the classifier will only succeed if it is capable of assigning unlearned words to test examples.

A relaxation on the obstacles of zero-shot learning is present in one-shot learning [39], where algorithms attempt to generalize from very few (1 to 5) examples of each class. This is a common circumstance in the field of image classification, where the cost of collecting and labeling data samples is high.

A classification of these problems according to the type of missing information can be found in Table III.3.

Table III.3.: Partial information problems according to the kind of absence in the training set.

Trait	Problem types
Presence of unlabeled instances	Semi-supervised [35], Positive-unlabeled [37]
No representation of some classes	One-class [36], Positive-unlabeled [37], Zero-shot [38]
Scarce representation of some classes	One-shot [39]

Prediction of structured data

The nonstandard variations described in this work generalize traditional supervised problems where the predicted output is at most a vector whose components take values in either a finite set or $\mathbb{R}$. Further generalizations are possible if other kinds of structures are allowed. For example, the target may take the form of an ordered sequence or a tree. In this case, the problem usually enters the scope of structured prediction [123], a generalization of supervised learning where methods must build structured data associated to input instances.

A particular case of supervised problem which can be seen under the umbrella of structured prediction is learning to rank [32], which does not involve a label space as such. Instead, training consists in learning from a set of feature vectors with a series of preferences among them, that is, a partial or total order in the training set. During testing a set of feature vectors is provided and the desired output is a ranking (with a predefined number of relevance levels, allowing ties) or a sorting (simply an ordering of the instances). This problem differs from OR in that individual classifications are usually meaningless: only relative distances among ranked instances matter.

III.7. Conclusions

Traditional supervised learning comprises two well known problems in machine learning: classification and regression. However, the multitude of applications which do not strictly fit the structure of the standard versions of those problems have favored the development of alternative versions which are more flexible and allow the analysis of more complex situations.

In this work an overview of nonstandard variations of supervised learning problems has been presented. A novel taxonomy under

several criteria has described relationships among these variations, where the main differentiating properties are multiplicity of inputs, multiplicity of outputs, presence of order relations and constraints, and partial information. Afterwards, common methods for tackling these problems have been outlined and their main applications have been mentioned as well. Finally, some additional variants which were left out of the scope of the previous analysis have been introduced as well.

Design of novel algorithms for nonstandard supervised tasks is scarcer than adaptations and transformations, but there exist some approximations and even more open possibilities for tackling these from classical algorithmic perspectives, such as probabilistic and heuristic methods, information theory and linear algebra, among others.

Acknowledgments

D. Charte is supported by the Spanish Ministry of Science, Innovation and Universities under the FPU National Program (Ref. FPU17/04069). This work has been partially supported by projects TIN2017-89517-P (FEDER Founds) of the Spanish Ministry of Economy and Competitiveness and TIN2015-68454-R of the Spanish Ministry of Science, Innovation and Universities.

References

[1] Tom M. Mitchell. *Machine learning*. McGraw Hill series in computer science. McGraw-Hill, 1997 (cit. on p. 1).

[2] Stephen Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman & Hall, 2014 (cit. on p. 2).

[3] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013 (cit. on pp. 2, 12).

[4] Anil K Jain, Robert PW Duin, and Jianchang Mao. "Statistical pattern recognition: A review". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.1 (2000), pp. 4–37 (cit. on p. 2).

[5] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012 (cit. on p. 2).

[6] David MJ Tax and Robert PW Duin. "Using two-class classifiers for multiclass classification". In: *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. Vol. 2. IEEE. 2002, pp. 124–127 (cit. on p. 2).

[7] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer New York, 2013 (cit. on p. 2).

[8] Alex J Smola and Bernhard Schölkopf. "On a kernel-based method for pattern recognition, regression, approximation, and operator inversion". In: *Algorithmica* 22.1-2 (1998), pp. 211–231 (cit. on pp. 2, 12).

[9] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. *Feature Selection for High-Dimensional Data*. Cham: Springer International Publishing, 2015 (cit. on p. 3).

[10] Alberto Fernández et al. *Learning from Imbalanced Data Sets*. Springer International Publishing, 2018 (cit. on p. 3).

[11] Bartosz Krawczyk. "Learning from imbalanced data: open challenges and future directions". In: *Progress in Artificial Intelligence* 5.4 (Nov. 2016), pp. 221–232. DOI: 10.1007/s13748-016-0094-0 (cit. on p. 3).

[12] Joao Gama. *Knowledge discovery from data streams*. Chapman and Hall/CRC, 2010 (cit. on p. 3).

[13] Jonathan A Silva et al. "Data stream clustering: A survey". In: *ACM Computing Surveys (CSUR)* 46.1 (2013), p. 13 (cit. on p. 3).

[14] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018 (cit. on p. 3).

[15] Jerónimo Hernández-González, Iñaki Inza, and Jose A. Lozano. "Weak supervision and other non-standard classification problems: A taxonomy". In: *Pattern Recognition Letters* 69 (2016), pp. 49–55. DOI: 10.1016/j.patrec.2015.10.008 (cit. on p. 3).

[16] Christopher KI Williams and David Barber. "Bayesian classification with Gaussian processes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.12 (1998), pp. 1342–1351 (cit. on p. 4).

[17] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012 (cit. on p. 4).

[18] Francisco Herrera et al. *Multiple instance learning: foundations and algorithms*. Springer, 2016 (cit. on pp. 4, 12).

[19] James Foulds and Eibe Frank. "A review of multi-instance learning assumptions". In: *The Knowledge Engineering Review* 25.1 (2010), pp. 1–25. DOI: 10.1017/S026988890999035X (cit. on p. 4).

[20] Jing Zhao et al. "Multi-view learning overview: Recent progress and new challenges". In: *Information Fusion* 38 (2017), pp. 43–54. DOI: 10.1016/j.inffus.2017.02.007 (cit. on pp. 5, 12).

[21] Francisco Herrera et al. *Multilabel classification*. Springer, 2016 (cit. on pp. 5, 12).

[22] Eva Gibaja and Sebastián Ventura. "A tutorial on multilabel learning". In: *ACM Computing Surveys (CSUR)* 47.3 (2015), p. 52. DOI: 10.1145/2716262 (cit. on p. 5).

[23] Francisco Charte et al. "Dealing with difficult minority labels in imbalanced mutilabel data sets". In: *Neurocomputing* (2017). DOI: 10.1016/j.neucom.2016.08.158 (cit. on p. 5).

[24] Carlos N Silla and Alex A Freitas. "A survey of hierarchical classification across different application domains". In: *Data Mining and Knowledge Discovery* 22.1-2 (2011), pp. 31–72 (cit. on p. 5).

[25] Hagit Shatkay et al. "Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users". In: *Bioinformatics* 24.18 (2008), pp. 2086–2093. DOI: 10.1093/bioinformatics/btn381 (cit. on pp. 6, 12, 13, 15–17).

[26] Xin Geng. "Label distribution learning". In: *IEEE Transactions on Knowledge and Data Engineering* 28.7 (2016), pp. 1734–1748. DOI: 10.1109/TKDE.2016.2545658 (cit. on pp. 6, 13, 15, 16).

[27] Pedro L López-Cruz, Concha Bielza, and Pedro Larrañaga. "Learning conditional linear Gaussian classifiers with probabilistic class labels". In: *Conference of the Spanish Association for Artificial Intelligence*. Springer. 2013, pp. 139–148. DOI: 10.1007/978-3-642-40643-0_15 (cit. on p. 6).

[28] Eyke Hüllermeier et al. "Label ranking by learning pairwise preferences". In: *Artificial Intelligence* 172.16-17 (2008), pp. 1897–1916. DOI: 10.1016/j.artint.2008.08.002 (cit. on pp. 6, 13, 16, 17).

[29] Shankar Vembu and Thomas Gärtner. "Label ranking algorithms: A survey". In: *Preference learning*. Springer, 2010, pp. 45–64. DOI: 10.1007/978-3-642-14125-6_3 (cit. on pp. 6, 15, 16).

[30] Hanen Borchani et al. "A survey on multi-output regression". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.5 (2015), pp. 216–233. DOI: 10.1002/widm.1157 (cit. on pp. 7, 12).

[31] P. A. Gutiérrez et al. "Ordinal Regression Methods: Survey and Experimental Study". In: *IEEE Transactions on Knowledge and Data Engineering* 28.1 (2016), pp. 127–146. DOI: 10.1109/TKDE.2015.2457911 (cit. on pp. 7, 12, 14, 16).

[32] Chris Burges et al. "Learning to rank using gradient descent". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 89–96. DOI: 10.1145/1102351.1102363 (cit. on pp. 7, 19).

[33] Pedro Antonio Gutiérrez and Salvador García. "Current prospects on ordinal and monotonic classification". In: *Progress in Artificial Intelligence* 5.3 (Aug. 2016), pp. 171–179. DOI: 10.1007/s13748-016-0088-y (cit. on pp. 8, 12).

[34] Richard E Barlow. *Statistical inference under order restrictions; the theory and application of isotonic regression*. Wiley, 1972 (cit. on pp. 8, 12).

[35] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. 1st. The MIT Press, 2010 (cit. on pp. 8, 18, 19).

[36] Mary M Moya, Mark W Koch, and Larry D Hostetler. "One-class classifier networks for target recognition applications". In: *NASA STI/Recon Technical Report N* 93 (1993) (cit. on pp. 8, 18, 19).

[37] Charles Elkan and Keith Noto. "Learning classifiers from only positive and unlabeled data". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 213–220. DOI: 10.1145/1401890.1401920 (cit. on pp. 8, 18, 19).

[38] Mark Palatucci et al. "Zero-shot learning with semantic output codes". In: *Advances in neural information processing systems*. 2009, pp. 1410–1418 (cit. on pp. 8, 18, 19).

[39] Li Fe-Fei et al. "A Bayesian approach to unsupervised one-shot learning of object categories". In: *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE. 2003, pp. 1134–1141. DOI: 10.1109/ICCV.2003.1238476 (cit. on pp. 8, 18, 19).

[40] Zhi-Hua Zhou et al. "Multi-instance multi-label learning". In: *Artificial Intelligence* 176.1 (2012), pp. 2291–2320. DOI: 10.1016/j.artint.2011.10.002 (cit. on pp. 8, 12).

[41] Mihai Surdeanu et al. “Multi-instance multi-label learning for relation extraction”. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics. 2012, pp. 455–465 (cit. on p. 8).

[42] Cam-Tu Nguyen et al. “Labeling Complicated Objects: Multi-View Multi-Instance Multi-Label Learning.” In: *AAAI*. 2014, pp. 2013–2019 (cit. on pp. 9, 12).

[43] Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. “Graded multilabel classification: The ordinal case”. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 223–230 (cit. on pp. 9, 12).

[44] Bin Wu et al. “Music emotion recognition by multi-label multi-layer multi-instance multi-view learning”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 117–126. DOI: 10.1145/2647868.2654904 (cit. on p. 9).

[45] Jaume Amores. “Multiple instance classification: Review, taxonomy and comparative study”. In: *Artificial Intelligence* 201 (2013), pp. 81–105. DOI: https://doi.org/10.1016/j.artint.2013.06.003 (cit. on pp. 12, 14, 16).

[46] Abhishek Kumar, Piyush Rai, and Hal Daume. “Co-regularized multi-view spectral clustering”. In: *Advances in neural information processing systems*. 2011, pp. 1413–1421 (cit. on p. 13).

[47] Kamalika Chaudhuri et al. “Multi-view clustering via canonical correlation analysis”. In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 129–136. DOI: 10.1145/1553374.1553391 (cit. on pp. 13, 16).

[48] Min-Ling Zhang and Zhi-Hua Zhou. “A review on multi-label learning algorithms”. In: *IEEE transactions on knowledge and data engineering* 26.8 (2014), pp. 1819–1837. DOI: 10.1109/TKDE.2013.39 (cit. on pp. 13, 15, 16).

[49] Grigorios Tsoumakas and Ioannis Vlahavas. “Random k-labelsets: An ensemble method for multilabel classification”. In: *European conference on machine learning*. Springer. 2007, pp. 406–417. DOI: 10.1007/978-3-540-74958-5_38 (cit. on p. 13).

[50] Jesse Read et al. “Classifier chains for multi-label classification”. In: *Machine learning* 85.3 (2011), p. 333. DOI: 10.1007/s10994-011-5256-5 (cit. on pp. 13, 16).

[51] Johannes Fürnkranz et al. “Multilabel classification via calibrated label ranking”. In: *Machine learning* 73.2 (2008), pp. 133–153. DOI: 10.1007/s10994-008-5064-8 (cit. on p. 13).

[52] Fengxia Pan. *Multi-dimensional fragment classification in biomedical text*. Queen’s University, 2006 (cit. on pp. 13, 15, 16).

[53] Sariel Har-Peled, Dan Roth, and Dav Zimak. “Constraint classification for multiclass classification and ranking”. In: *Advances in neural information processing systems*. 2003, pp. 809–816 (cit. on pp. 13, 16).

[54] Nils J Nilsson. *Learning machines: foundations of trainable pattern-classifying systems*. McGraw-Hill, 1965 (cit. on p. 13).

[55] Eleftherios Spyromitros-Xioufis et al. “Multi-label classification methods for multi-target regression”. In: *arXiv preprint arXiv* 1211 (2012) (cit. on pp. 14, 16).

[56] Wei Zhang et al. “Multi-output LS-SVR machine in extended feature space”. In: *Computational Intelligence for Measurement Systems and Applications (CIMSA), 2012 IEEE International Conference on*. IEEE. 2012, pp. 130–134. DOI: 10.1109/CIMSA.2012.6269600 (cit. on pp. 14, 16).

[57] Eibe Frank and Mark Hall. "A simple approach to ordinal classification". In: *European Conference on Machine Learning*. Springer. 2001, pp. 145–156. DOI: 10.1007/3-540-44795-4_13 (cit. on pp. 14, 16).

[58] W. Kotlowski and R. Slowinski. "On Nonparametric Ordinal Classification with Monotonicity Constraints". In: *IEEE Transactions on Knowledge and Data Engineering* 25.11 (2013), pp. 2576–2589. DOI: 10.1109/TKDE.2012.204 (cit. on pp. 14, 16).

[59] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. "Support vector machines for multiple-instance learning". In: *Advances in neural information processing systems*. 2003, pp. 577–584 (cit. on pp. 14, 16).

[60] Jan Ramon and Luc De Raedt. "Multi instance neural networks". In: *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*. 2000, pp. 53–60 (cit. on pp. 14, 16).

[61] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial intelligence* 89.1-2 (1997), pp. 31–71. DOI: 10.1016/S0004-3702(96)00034-3 (cit. on pp. 14, 16).

[62] Oded Maron and Tomás Lozano-Pérez. "A framework for multiple-instance learning". In: *Advances in neural information processing systems*. 1998, pp. 570–576 (cit. on p. 14).

[63] Jun Wang and Jean-Daniel Zucker. "Solving Multiple-Instance Problem: a Lazy Learning Approach". In: *International Conference on Machine Learning*. Morgan Kaufmann Publishers. 2000, pp. 1119–1126 (cit. on pp. 14, 16).

[64] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. "Multi-instance learning by treating instances as non-iid samples". In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 1249–1256. DOI: 10.1145/1553374.1553534 (cit. on pp. 14, 16).

[65] Jason Farquhar et al. "Two view learning: SVM-2K, theory and practice". In: *Advances in neural information processing systems*. 2006, pp. 355–362 (cit. on pp. 15, 16).

[66] Qiaona Chen and Shiliang Sun. "Hierarchical multi-view fisher discriminant analysis". In: *International Conference on Neural Information Processing*. Springer. 2009, pp. 289–298. DOI: 10.1007/978-3-642-10684-2_32 (cit. on pp. 15, 16).

[67] Min-Ling Zhang and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning". In: *Pattern recognition* 40.7 (2007), pp. 2038–2048. DOI: 10.1016/j.patcog.2006.12.019 (cit. on pp. 15, 16).

[68] Amanda Clare and Ross D King. "Knowledge discovery in multi-label phenotype data". In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer. 2001, pp. 42–53. DOI: 10.1007/3-540-44794-6_4 (cit. on pp. 15, 16).

[69] André Elisseeff and Jason Weston. "A kernel method for multi-labelled classification". In: *Advances in neural information processing systems*. 2002, pp. 681–687 (cit. on pp. 15, 16).

[70] Fadi A Thabtah, Peter Cowling, and Yonghong Peng. "MMAC: A new multi-class, multi-label associative classification approach". In: *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE. 2004, pp. 217–224. DOI: 10.1109/ICDM.2004.10117 (cit. on pp. 15, 16).

[71] Jose M Moyano et al. "Review of ensembles of multi-label classifiers: Models, experimental study and prospects". In: *Information Fusion* 44 (2018), pp. 33–45. DOI: 10.1016/j.inffus.2017.12.001 (cit. on pp. 15, 16).

[72] Concha Bielza, Guangdi Li, and Pedro Larranaga. "Multi-dimensional classification with Bayesian networks". In: *International Journal of Approximate Reasoning* 52.6 (2011), pp. 705–727 (cit. on pp. 15–17).

[73] Peter R De Waal and Linda C Van Der Gaag. "Inference and learning in multi-dimensional Bayesian network classifiers". In: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer. 2007, pp. 501–511. DOI: 10.1007/978-3-540-75256-1_45 (cit. on pp. 15–17).

[74] Ofer Dekel, Yoram Singer, and Christopher D Manning. "Log-linear models for label ranking". In: *Advances in neural information processing systems*. 2004, pp. 497–504 (cit. on pp. 15, 16).

[75] Shai Shalev-Shwartz and Yoram Singer. "A unified algorithmic approach for efficient online label ranking". In: *Artificial Intelligence and Statistics*. 2007, pp. 452–459 (cit. on pp. 15, 16).

[76] Alan Julian Izenman. "Reduced-rank regression for the multivariate linear model". In: *Journal of multivariate analysis* 5.2 (1975), pp. 248–264. DOI: 10.1016/0047-259X(75)90042-1 (cit. on pp. 15, 16).

[77] A Van Der Merwe and JV Zidek. "Multivariate regression analysis and canonical variates". In: *Canadian Journal of Statistics* 8.1 (1980), pp. 27–39. DOI: 10.2307/3314667 (cit. on pp. 15, 16).

[78] Emmanuel Vazquez and Eric Walter. "Multi-output support vector regression". In: *13th IFAC Symposium on System Identification*. Citeseer. 2003, pp. 1820–1825 (cit. on pp. 15, 16).

[79] Matilde Sánchez-Fernández et al. "SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems". In: *IEEE transactions on signal processing* 52.8 (2004), pp. 2298–2307. DOI: 10.1109/TSP.2004.831028 (cit. on pp. 15, 16).

[80] Charles A Micchelli and Massimiliano Pontil. "On learning vector-valued functions". In: *Neural computation* 17.1 (2005), pp. 177–204. DOI: 10.1162/0899766052530802 (cit. on pp. 15, 16).

[81] Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. "Kernels for Vector-Valued Functions: A Review". In: *Foundations and Trends in Machine Learning*. Now Publishers, 2012. DOI: 10.1561/2200000036 (cit. on pp. 15, 16).

[82] Glenn De'Ath. "Multivariate regression trees: a new technique for modeling species–environment relationships". In: *Ecology* 83.4 (2002), pp. 1105–1117. DOI: 10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2 (cit. on pp. 15, 16).

[83] Dragi Kocev et al. "Tree ensembles for predicting structured outputs". In: *Pattern Recognition* 46.3 (2013), pp. 817–833. DOI: 10.1016/j.patcog.2012.09.023 (cit. on pp. 15, 16).

[84] Mario Costa. "Probabilistic interpretation of feedforward network outputs, with relationships to statistical prediction of ordinal quantities". In: *International journal of neural systems* 7.05 (1996), pp. 627–637. DOI: 10.1142/S0129065796000610 (cit. on pp. 15, 16).

[85] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. "A neural network approach to ordinal regression". In: *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE. 2008, pp. 1279–1284. DOI: 10.1109/IJCNN.2008.4633963 (cit. on pp. 15, 16).

[86] Wan-Yu Deng et al. "Ordinal extreme learning machine". In: *Neurocomputing* 74.1-3 (2010), pp. 447–456. DOI: 10.1016/j.neucom.2010.08.022 (cit. on pp. 15, 16).

[87] Javier Sánchez-Monedero, Pedro Antonio Gutiérrez, and Cesar Hervás-Martínez. "Evolutionary ordinal extreme learning machine". In: *International Conference on Hybrid Artificial Intelligence Systems*. Springer. 2013, pp. 500–509. DOI: 10.1007/978-3-642-40846-5_50 (cit. on pp. 15, 16).

[88] Jaime S Cardoso and Ricardo Sousa. "Classification models with global constraints for ordinal data". In: *2010 Ninth International Conference on Machine Learning and Applications*. IEEE. 2010, pp. 71–77. DOI: 10.1109/ICMLA.2010.18 (cit. on pp. 15, 16).

[89] Wei Chu and Zoubin Ghahramani. "Gaussian processes for ordinal regression". In: *Journal of machine learning research* 6.Jul (2005), pp. 1019–1041 (cit. on pp. 15, 16).

[90] Hsuan-Tien Lin and Ling Li. "Combining ordinal preferences by boosting". In: *Proceedings ECML/PKDD 2009 Workshop on Preference Learning*. 2009, pp. 69–83 (cit. on pp. 15, 16).

[91] Wouter Duivesteijn and Ad Feelders. "Nearest neighbour classification with monotonicity constraints". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2008, pp. 301–316. DOI: 10.1007/978-3-540-87479-9_38 (cit. on p. 16).

[92] Rob Potharst and Adrianus Johannes Feelders. "Classification trees for problems with monotonicity constraints". In: *ACM SIGKDD Explorations Newsletter* 4.1 (2002), pp. 1–10. DOI: 10.1145/568574.568577 (cit. on pp. 16, 18).

[93] Krzysztof Dembczyński, Wojciech Kotłowski, and Roman Słowiński. "Ensemble of decision rules for ordinal classification with monotonicity constraints". In: *International Conference on Rough Sets and Knowledge Technology*. Springer. 2008, pp. 260–267. DOI: 10.1007/978-3-540-79721-0_38 (cit. on p. 16).

[94] Jerzy Błaszczyński, Roman Słowiński, and Marcin Szelag. "Sequential covering rule induction algorithm for variable consistency rough set approaches". In: *Information Sciences* 181.5 (2011), pp. 987–1002. DOI: 10.1016/j.ins.2010.10.030 (cit. on p. 16).

[95] Joseph Sill. "Monotonic networks". In: *Advances in neural information processing systems*. 1998, pp. 661–667 (cit. on p. 16).

[96] Sotiris Kotsiantis, Dimitris Kanellopoulos, and Vasilis Tampakas. "Financial application of multi-instance learning: two greek case studies". In: *Journal of Convergence Information Technology* 5.8 (2010), pp. 42–53 (cit. on p. 16).

[97] Massih Amini, Nicolas Usunier, and Cyril Goutte. "Learning from multiple partially observed views-an application to multilingual text categorization". In: *Advances in neural information processing systems*. 2009, pp. 28–36 (cit. on p. 17).

[98] Stan Z Li et al. "Statistical learning of multi-view face detection". In: *European Conference on Computer Vision*. Springer. 2002, pp. 67–81. DOI: 10.1007/3-540-47979-1_5 (cit. on p. 17).

[99] Sinno Jialin Pan et al. "Adaptive localization in a dynamic WiFi environment through multi-view learning". In: *AAAI*. 2007, pp. 1108–1113 (cit. on p. 17).

[100] Shiliang Sun and Guoqing Chao. "Multi-View Maximum Entropy Discrimination." In: *IJCAI*. 2013, pp. 1706–1712 (cit. on p. 17).

[101] Grigorios Tzortzis and Aristidis Likas. "Kernel-based weighted multi-view clustering". In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE. 2012, pp. 675–684. DOI: 10.1109/ICDM.2012.43 (cit. on p. 17).

[102] I. Katakis, G. Tsoumakas, and I. Vlahavas. "Multilabel Text Classification for Automated Tag Suggestion". In: *Proc. ECML PKDD08 Discovery Challenge, Antwerp, Belgium*. 2008, pp. 75–83 (cit. on p. 17).

[103] M. Boutell et al. "Learning multi-label scene classification". In: *Pattern Recognition* 37.9 (2004), pp. 1757–1771. DOI: 10.1016/j.patcog.2004.03.009 (cit. on p. 17).

[104] Francisco Charte et al. "QUINTA: A question tagging assistant to improve the answering ratio in electronic forums". In: *EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), IEEE*. Sept. 2015, pp. 1–6. DOI: 10.1109/EUROCON.2015.7313677 (cit. on p. 17).

[105] S. Diplaris et al. "Protein Classification with Multiple Algorithms". In: *Proc. 10th Panhellenic Conference on Informatics, Volos, Greece, PCI05*. 2005, pp. 448–456. DOI: 10.1007/11573036_42 (cit. on p. 17).

[106] Ricardo Sousa and João Gama. "Multi-label classification from high-speed data streams with adaptive model rules and random rules". In: *Progress in Artificial Intelligence* 7.3 (Sept. 2018), pp. 177–187. DOI: 10.1007/s13748-018-0142-z (cit. on p. 17).

[107] Khalil Laghmari, Christophe Marsala, and Mohammed Ramdani. "An adapted incremental graded multi-label classification model for recommendation systems". In: *Progress in Artificial Intelligence* 7.1 (Mar. 2018), pp. 15–29. DOI: 10.1007/s13748-017-0133-5 (cit. on p. 17).

[108] Johannes Fürnkranz et al. "Multilabel classification via calibrated label ranking". In: *Machine learning* 73.2 (2008), pp. 133–153. DOI: 10.1007/s10994-008-5064-8 (cit. on p. 17).

[109] Michael B Eisen et al. "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868 (cit. on p. 17).

[110] Michael Lyons et al. "Coding facial expressions with gabor wavelets". In: *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE. 1998, pp. 200–205. DOI: 10.1109/AFGR.1998.670949 (cit. on p. 17).

[111] Dragi Kocev et al. "Using single-and multi-target regression trees and ensembles to model a compound index of vegetation condition". In: *Ecological Modelling* 220.8 (2009), pp. 1159–1168. DOI: 10.1016/j.ecolmodel.2009.01.037 (cit. on p. 17).

[112] Damjan Kuznar, Martin Mozina, and Ivan Bratko. "Curve prediction with kernel regression". In: *Proceedings of the 1st Workshop on Learning from Multi-Label Data*. 2009, pp. 61–68 (cit. on p. 17).

[113] Devis Tuia et al. "Multioutput support vector regression for remote sensing biophysical parameter estimation". In: *IEEE Geoscience and Remote Sensing Letters* 8.4 (2011), pp. 804–808. DOI: 10.1109/LGRS.2011.2109934 (cit. on p. 17).

[114] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. "Feature selection for ordinal text classification". In: *Neural computation* 26.3 (2014), pp. 557–591. DOI: 10.1162/NECO_a_00558 (cit. on p. 17).

[115] Qing Tian, Songcan Chen, and Xiaoyang Tan. "Comparative study among three strategies of incorporating spatial structures to ordinal image regression". In: *Neurocomputing* 136 (2014), pp. 152–161. DOI: 10.1016/j.neucom.2014.01.017 (cit. on p. 17).

[116] Ralf Bender and Ulrich Grouven. "Ordinal logistic regression in medical research". In: *Journal of the Royal College of physicians of London* 31.5 (1997), pp. 546–551 (cit. on p. 17).

[117] Young S Kwon, Ingoo Han, and Kun Chang Lee. "Ordinal pairwise partitioning (OPP) approach to neural networks training in bond rating". In: *Intelligent Systems in Accounting, Finance & Management* 6.1 (1997), pp. 23–40. DOI: 10.1002/(SICI)1099-1174(199703)6:1<23::AID-ISAF113>3.0.CO;2-4 (cit. on p. 17).

[118] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. "Ordinal hyperplanes ranker with cost sensitivities for age estimation". In: *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*. IEEE. 2011, pp. 585–592. DOI: 10.1109/CVPR.2011.5995437 (cit. on p. 17).

[119] Salvatore Greco, Benedetto Matarazzo, and Roman Słowiński. "Rough set approach to customer satisfaction analysis". In: *International Conference on Rough Sets and Current Trends in Computing*. Springer. 2006, pp. 284–295. DOI: 10.1007/11908029_31 (cit. on p. 18).

[120] Salvatore Greco, Benedetto Matarazzo, and Roman Slowinski. "A new rough set approach to evaluation of bankruptcy risk". In: *Operational tools in the management of financial risks*. Springer, 1998, pp. 121–136. DOI: 10.1007/978-1-4615-5495-0_8 (cit. on p. 18).

[121] Young U Ryu, Ramaswamy Chandrasekaran, and Varghese S Jacob. "Breast cancer prediction using the isotonic separation technique". In: *European Journal of Operational Research* 181.2 (2007), pp. 842–854. DOI: 10.1016/j.ejor.2006.06.031 (cit. on p. 18).

[122] Bing Liu et al. "Building text classifiers using positive and unlabeled examples". In: *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE. 2003, pp. 179–186. DOI: 10.1109/ICDM.2003.1250918 (cit. on p. 18).

[123] Ben Taskar et al. "Learning structured prediction models: A large margin approach". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 896–903. DOI: 10.1145/1102351.1102464 (cit. on p. 19).