ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200238

General Supervision via Probabilistic Transformations

Santiago Mazuelas[1] and **Aritz Pérez**[2]

Abstract. Different types of training data have led to numerous schemes for supervised classification. Current learning techniques are tailored to one specific scheme and cannot handle general ensembles of training samples. This paper presents a unifying framework for supervised classification with general ensembles of training samples, and proposes the learning methodology of generalized robust risk minimization (GRRM). The paper shows how current and novel supervision schemes can be addressed under the proposed framework by representing the relationship between examples at prediction and training via probabilistic transformations. The results show that GRRM can handle different types of training samples in a unified manner, and enable new supervision schemes that aggregate general ensembles of training samples.

1 Introduction

Supervised classification uses training samples to choose a classification rule with small expected loss over variables at prediction (instance and label). Since the actual probability distribution of prediction variables is unknown, expected losses are evaluated with respect to a probability distribution obtained from training samples. Approaches based on empirical risk minimization (ERM) use the empirical distribution of training samples [25, 8] while approaches based on robust risk minimization (RRM) use a distribution with maximum entropy near the empirical distribution [9, 20, 2, 13].

In standard supervision, examples at training follow the same distribution as examples at prediction, while numerous non-standard supervision schemes have been proposed to exploit more general types of training samples. Current non-standard schemes consider: i) labels at training that are less precise than those at prediction; ii) instances at training that are more informative than those at prediction; iii) instances at training that are less informative than those at prediction; iv) examples at training that are from a different domain; v) examples at training with missing components; and vi) examples at training with multiple qualities and domains. Those schemes have been developed under different paradigms such as weak supervision, semi-supervision, privileged information, and domain adaptation (see specific current approaches and associated references in Sections 3 and 4).

The diverse range of supervision schemes described above can be particularly useful in practice. Schemes that use training examples from different domains or less precise than prediction examples can reduce training costs, while those that use training examples more precise than prediction examples can increase classification accuracies. Current techniques are tailored to one specific supervision scheme and there is a lack of a common methodology for supervised classification with general training data. As a consequence, it is currently not possible to adequately deal with cost/accuracy trade-offs and to seamlessly develop versatile algorithms. For instance, existing techniques can only handle scenarios with training samples in accordance with one of the specific cases described above, and cannot exploit general ensembles of training samples with assorted types and qualities. This paper presents a unifying framework for supervised classification with general ensembles of training samples, and proposes the learning methodology of generalized RRM (GRRM). Such framework is enabled by representing the relationship between examples at prediction and training stages via probabilistic transformations. The paper shows how current and novel supervision schemes can be addressed under the proposed framework. In particular, we show that GRRM can enable learning algorithms that aggregate general ensembles of training samples with different types.

[1] BCAM-Basque Center for Applied Mathematics and IKERBASQUE-Basque Foundation for Science, Spain, email: smazuelas@bcamath.org
[2] BCAM-Basque Center for Applied Mathematics, Spain, email: aperez@bcamath.org

2 Preliminaries

This section provides an overview of the supervised classification problem, recalls the notion of probabilistic transformation, and describes notations used in the rest of the paper. In particular, in the following, upright letters denote random variables (RVs); calligraphic upper case letters denote sets; $\mathbb{I}\{\cdot\}$ denotes the indicator function; $\mathbb{E}_{a\sim p}\{f(a)\}$ or just $\mathbb{E}_p\{f(a)\}$ denotes the expectation of function f over instantiations a that follow probability distribution p; and I denotes the identity transformation.

2.1 Supervised classification

A problem of supervised classification can be described by four objects $(\mathrm{Z}, \mathrm{D}, \mathcal{H}, L)$ representing variables at prediction, training data, classification rules, and miss-classification losses. Specifically, $\mathrm{Z} = (\mathrm{X}, \mathrm{Y})$ is an RV representing examples at prediction, X is called instance or attribute, and Y has finite support and is called label or class. D is an RV describing training data formed by the concatenation of training samples. For instance, in standard supervision each instantiation of D is $d = (z^{(1)}, z^{(2)}, \ldots, z^{(n)})$ where $z^{(i)}$ for $i = 1, 2, \ldots, n$ are independent instantiations of Z. The classification rules $\mathcal{H}$ are mappings from instances to labels, i.e., $h \in \mathcal{H}$, $h : \mathcal{X} \to \mathcal{Y}$. Finally, L is a function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $L(\hat{y}, y)$ quantifies the loss of predicting the label y by label $\hat{y}$, e.g., $L(\hat{y}, y) = \mathbb{I}\{y \neq \hat{y}\}$.

The goal of a learning algorithm for classification is to determine a rule $h \in \mathcal{H}$ with small expected loss (risk) under the probability distribution of Z, p, that is, to solve the optimization problem

$$\min_{h\in\mathcal{H}} \mathbb{E}_p\big\{L(h(x), y)\big\}. \tag{1}$$

Training data aids the learning problem in that it provides information regarding the probability distribution p.

Supervised learning based on ERM corresponds to solving (1) using the empirical distribution p_e of the training data d instead of p. The main drawback of ERM approach is over-fitting that is often addressed using regularization methods. Most techniques for regularization are based on structural ERM that considers subsets of classification rules with reduced complexity [25, 8]. Other complementary regularization techniques are based on RRM that considers uncertainty (ambiguity) sets $\mathcal{U}$ of probability distributions [9, 20, 2, 13]. Specifically, the classification rule in such techniques is obtained by minimizing the maximum expected loss over the uncertainty set, i.e., solving

$$\min_{h\in\mathcal{H}} \max_{q\in\mathcal{U}} \mathbb{E}_q\big\{L(h(x),y)\big\}. \tag{2}$$

The uncertainty set $\mathcal{U}$ is formed by distributions close to the empirical distribution, where the closeness between distributions in $\mathcal{Z}$ is quantified by a discrepancy function ψ, hence

$$\mathcal{U} = \{q \in \Delta(\mathcal{Z}) :\ \psi(q, p_e) < \varepsilon\}$$

with $\Delta(\mathcal{Z})$ the set of probability distributions supported in $\mathcal{Z}$. For instance, the uncertainty sets used in [20, 13] correspond to consider as $\psi(q_1, q_2)$ the Wasserstein (transportation) distance between q_1 and q_2, while those used in [9] correspond to

$$\psi(q_1, q_2) = \|\mathbb{E}_{q_1}\{t(z)\} - \mathbb{E}_{q_2}\{t(z)\}\|_2 \tag{3}$$

for q_1 and q_2 distributions with the same marginal over $\mathcal{X}$, and $t(\cdot)$ a statistic over $\mathcal{Z}$.

For each distribution $q \in \Delta(\mathcal{Z})$, the minimum expected loss defines an entropy function as $H(q) = \min_{h\in\mathcal{H}} \mathbb{E}_q\big\{L(h(x),y)\big\}$ [11]. For instance, if $L(\hat{y}, y) = \mathbb{I}(y \neq \hat{y})$ and $\mathcal{H}$ contains the posterior Bayes rule, the entropy is given by

$$H(q) = \mathbb{E}_q\left\{1 - \max_{y\in\mathcal{Y}} q(y|x)\right\} = 1 - \int \max_{y\in\mathcal{Y}} q(x,y)dx \tag{4}$$

where $q(y|x)$ denotes the conditional distribution of Y given X for q. Under mild regularity conditions [9, 11], the minimax solution of (2) coincides with its maximin solution. Therefore, RRM methods solve (2) using as surrogate of q the distribution q^* that maximizes the associated entropy near the empirical distribution, i.e.,

$$q^* = \arg\min_q \psi(q, p_e) - \lambda H(q) \tag{5}$$

for a regularization parameter λ. Both ERM and RRM strategies are often equivalent [3]. However, the empirical distribution of non-standard training samples is often not adequate to assess the uncertainty about prediction variables (see Section 3.2 below), and in this paper we extend the RRM approach for non-standard supervision.

2.2 Probability distributions and probabilistic transformations

Probabilistic transformations, also known as Markov transitions or just transitions [1, 24], are a generalization of the concept of deterministic transformation and allow to represent random and uncertain processes. In the following, for each support set $\mathcal{V}$, a probability distribution $q \in \Delta(\mathcal{V})$ is given by a function on $\mathcal{V}$, e.g., density function or probability mass function.[3]

[3] We consider RVs with probability measures dominated by a base measure. More general scenarios can be analogously treated by requiring certain measure-theoretic regularity conditions such as Borel probability measures and Polish spaces, see for instance [1, 11].

Definition 1. A probabilistic transformation is a linear map that transforms probability distributions into probability distributions. For support sets $\mathcal{V}$ and $\mathcal{W}$, we denote by $\Delta(\mathcal{V}, \mathcal{W})$ the set of probabilistic transformations T with $T(q) \in \Delta(\mathcal{W})$ for $q \in \Delta(\mathcal{V})$.

If $\mathcal{V}$ and $\mathcal{W}$ have n and m elements, respectively, a probabilistic transformation in $\Delta(\mathcal{V}, \mathcal{W})$ is given by a $m \times n$ column-stochastic Markov transition matrix K; then $T(q) = r$ given by $r(w) = \sum_{v\in\mathcal{V}} K(w, v)q(v)$, with $K(w, v)$ the matrix component in row $w \in \mathcal{W}$ and column $v \in \mathcal{V}$. Analogously, for infinite sets, a probabilistic transformation in $\Delta(\mathcal{V}, \mathcal{W})$ is given by a function $K(w, v)$ called Markov transition kernel, then $T(q) = r$ given by $r(w) = \int_{\mathcal{V}} K(w, v)q(v)dv$. Simple examples of probabilistic transformations are deterministic and set-valued functions $f : \mathcal{V} \to \mathcal{W}$ in which the image of a distribution supported in a single point v is a uniform probability distribution with support $f(v)$. In addition, the conditional distribution of an RV W conditioned on an RV V provides a probabilistic transformation denoted $T_{\mathrm{W|V}}$ that maps the probability distribution of V to that of W.

Probabilistic transformations can be composed in series and in parallel. For instance, if $T_1 \in \Delta(\mathcal{V}_1, \mathcal{W}_1)$ and $T_2 \in \Delta(\mathcal{V}_2, \mathcal{W}_2)$ are given by Markov transitions kernels $K_1(w_1, v_1)$ and $K_2(w_2, v_2)$, respectively, the parallel composition of T_1 and T_2 denoted $T_1 \otimes T_2 \in \Delta(\mathcal{V}_1 \times \mathcal{V}_2, \mathcal{W}_1 \times \mathcal{W}_2)$ is given by the Markov transition kernel $K_1(w_1, v_1)K_2(w_2, v_2)$. For finite support sets, composition in series and parallel corresponds to matrix multiplication and Kronecker product, respectively.

3 Supervision with non-standard training samples

In this section we consider non-standard supervision cases in which examples at prediction and training are instantiations of two possibly different RVs Z and $\tilde{\mathrm{Z}}$, that is, training samples are $d = (\tilde{z}^{(1)}, \tilde{z}^{(2)}, \ldots, \tilde{z}^{(n)})$ where $\tilde{z}^{(i)}$ for $i = 1, 2, \ldots, n$ are independent instantiations of $\tilde{\mathrm{Z}}$. Several current supervision schemes use non-standard training data such as:

- Noisy labels [16, 14]: labels at prediction and training take the same categorical values, but training labels are affected by errors.
- Multiple labels [12]: labels at prediction are single categorical values and labels at training are sets of categorical values.
- Weak multi-labels [23]: labels at prediction are sets of categorical values and labels at training are partial sets of categorical values.
- Privileged information [18]: instances at training stage have more components than those at prediction.
- Prediction stage (PS) corrupted instances [7]: instances at prediction are corrupted by noise.
- Training stage (TS) corrupted instances [21]: instances at training are corrupted by noise.
- Representation based (RB) domain adaptation [5]: examples at prediction and training belong to different domains that share a common representation.
- Covariate shift [22]: variables at prediction and training share the same conditional distribution of labels given instances, but instances at prediction and training have different marginal distributions.

In the following we present a unifying framework for non-standard supervision, and describe how current and novel schemes can be addressed under such framework.

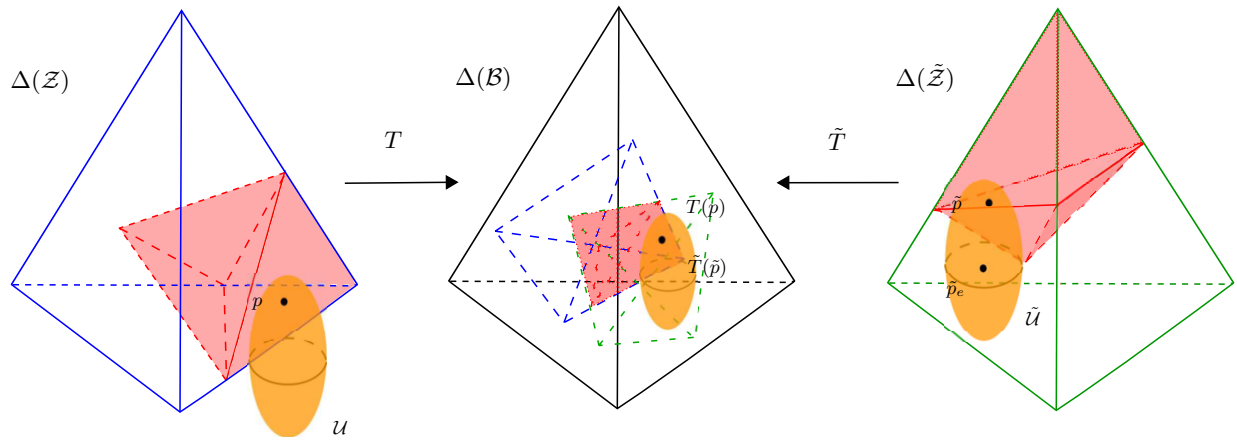

Figure 1. The relationship between prediction and training variables imposes structural constraints for feasible distributions (red polyhedra), and enables to use information from training samples as uncertainty sets (orange ellipsoids).

3.1 Unifying framework for non-standard supervision

Let $\mathcal{B}$ be a support set, and $T \in \Delta(\mathcal{Z}, \mathcal{B})$ and $\tilde{T} \in \Delta(\tilde{\mathcal{Z}}, \mathcal{B})$ be probabilistic transformations such that $T(p) = \tilde{T}(\tilde{p})$ for p and $\tilde{p}$ the distributions of Z and $\tilde{\mathrm{Z}}$, respectively. $T(p) = \tilde{T}(\tilde{p})$ is the probability distribution of an RV B that we call *bridge* since it serves to extract information for Z from training samples in $\tilde{\mathrm{Z}}$. For instance, if prediction examples are affected by noisy instances and training examples are affected by noisy labels, a variable composed by noisy instances and noisy labels can serve as a bridge to extract the information in training samples (see third case study in Section 5). Probabilistic transformations T and $\tilde{T}$ represent the relationship between prediction and training variables, impose structural constrains into the distributions considered, and allow to extract the information in non-standard training data as follows (see Fig. 1). Feasible distributions $\mathcal{F} \subset \Delta(\mathcal{Z})$ and $\tilde{\mathcal{F}} \subset \Delta(\tilde{\mathcal{Z}})$ are

$$\mathcal{F} = \{q \in \Delta(\mathcal{Z}) : \exists\, \tilde{q} \in \Delta(\tilde{\mathcal{Z}}),\ T(q) = \tilde{T}(\tilde{q})\}$$

$$\tilde{\mathcal{F}} = \{\tilde{q} \in \Delta(\tilde{\mathcal{Z}}) : \exists\, q \in \Delta(\mathcal{Z}),\ \tilde{T}(\tilde{q}) = T(q)\}$$

Note that feasibility is a necessary condition to be the actual distribution of Z or $\tilde{\mathrm{Z}}$. One consequence of the above is that ERM approach is inadequate in these settings since the empirical distribution of training samples is often not feasible (see discussion for Equation (8) in Section 3.2 below).

The above probabilistic transformations also allow to define uncertainty sets $\mathcal{U} \subset \Delta(\mathcal{Z})$ as

$$\mathcal{U} = \{q \in \Delta(\mathcal{Z}) : \psi(T(q), \tilde{T}(\tilde{p}_e)) < \varepsilon\}$$

where ψ is a discrepancy function in $\Delta(\mathcal{B})$ and $\tilde{p}_e$ is the empirical distribution in $\Delta(\tilde{\mathcal{Z}})$ of training samples. Therefore, learning from non-standard training data can be approached analogously to RRM, substituting optimization in (5) by

$$\min_{q \in \mathcal{F}} \psi\big(T(q), \tilde{T}(\tilde{p}_e)\big) - \lambda H(q) \tag{6}$$

where $\lambda > 0$ is a regularization parameter. We call GRRM the approach given by using (6) above instead of (5). Note that it reduces to RRM in the case of standard supervision, i.e., $\mathrm{Z} = \tilde{\mathrm{Z}}$, but allows also to use non-standard training data via the structural constraints and uncertainty sets given by the probabilistic transformations T and $\tilde{T}$.

The implementation complexity of GRRM is also similar to that of RRM since both can be enabled by solving a convex optimization problem and their main difference lies on how the uncertainty set $\mathcal{U}$ is defined (by means of $\psi(T(q), \tilde{T}(\tilde{p}_e))$ instead of $\psi(q, p_e)$). Therefore, efficient implementations of GRRM can be devised similarly as for RRM, for instance by exploiting equivalent reformulations based on convex duality [9, 20, 2]. The determination of transformations T and $\tilde{T}$ in practice requires certain knowledge about the relationship between prediction and training variables, and possibly to estimate certain parameters similarly to current techniques, e.g., label noise probabilities [16, 14]. This requirement is to be expected and unavoidable since non-standard supervision uses information from training variables that is used for prediction variables. Note that in most scenarios, such as those described in Tables 1 and 2 below, the knowledge required to determine transformations T and $\tilde{T}$ is quite modest since the same transformations can be used with independence of the actual probability distributions of prediction and training variables.

3.2 Different non-standard supervision schemes under the proposed framework

Table 1 shows how different current supervision schemes can be addressed under the proposed framework, and how the probabilistic transformations T and $\tilde{T}$ represent the relationship between prediction and training variables. In certain supervision schemes, such as noisy labels, multiple lables, and TS corrupted intances, examples at training stage are less precise than those at prediction. Then, we can take $\mathrm{B} = \tilde{\mathrm{Z}}$ and $T \in \Delta(\mathcal{Z}, \tilde{\mathcal{Z}})$ the probabilistic transformation corresponding to the conditional distribution of training variables given prediction variables. In other schemes, such as privileged information and PS corrupted instances, examples at training stage are more precise than those at prediction. Then, we can take $\mathrm{B} = \mathrm{Z}$ and $\tilde{T} \in \Delta(\tilde{\mathcal{Z}}, \mathcal{Z})$ the probabilistic transformation corresponding to the conditional distribution of prediction variables given training variables. Yet in other schemes, such as RB domain adaptation, examples at prediction and training stages are not related by being more or less precise but can be related through an instances' representation. Then, we can take B as such common representation and

Table 1. Current non-standard supervision schemes.

Supervision scheme	Prediction $\mathrm{Z} = (\mathrm{X}, \mathrm{Y})$ vs training $\tilde{\mathrm{Z}} = (\tilde{\mathrm{X}}, \tilde{\mathrm{Y}})$	Bridge B	Prob. transformations
Noisy labels Multiple labels Weak multi-labels	$\mathrm{X} = \tilde{\mathrm{X}}$ $\tilde{y}$ noisy $\mathrm{X} = \tilde{\mathrm{X}}$ $\tilde{y}$ set, $y \in \tilde{y}$ $\mathrm{X} = \tilde{\mathrm{X}}$ $y, \tilde{y}$ sets, $\tilde{y} \subset y$	$\tilde{\mathrm{Z}}$	$T = I \otimes T_{\tilde{\mathrm{Y}}\mid\mathrm{Y}}$ $\tilde{T} = I$
Privileged information	$\tilde{\mathrm{X}} = (\mathrm{X}, \mathrm{X}^{\text{priv}})$	Z	$T = I$ $\tilde{T} = T_{\mathrm{X}\mid\tilde{\mathrm{X}}} \otimes I$
PS corrupted instances	x noisy		
TS corrupted instances	$\tilde{x}$ noisy	$\tilde{\mathrm{X}}$	$T = T_{\tilde{\mathrm{X}}\mid\mathrm{X}} \otimes I$ $\tilde{T} = I$
RB domain adaptation	$\mathcal{Y} = \tilde{\mathcal{Y}},\ \mathrm{Y} \neq \tilde{\mathrm{Y}}$	General	$T = \tilde{T} = T_{\mathrm{B}\mid\mathrm{Z}}$

$\tilde{T} = T \in \Delta(\mathcal{Z}, \mathcal{B})$ the probabilistic transformation corresponding to the function mapping instances to their representation.

The proposed framework can offer a common methodology for learning using non-standard training data based on GRRM that uses distribution q^* solving (6) instead of p in (1). In addition, such framework can bring new insights for the design of algorithms for supervised classification. For instance, certain existing approaches for noisy labels [16, 24] first transform loss functions in $\mathcal{Z}$ into loss functions in $\tilde{\mathcal{Z}}$ and then use the ERM approach in $\tilde{\mathcal{Z}}$. However, the empirical distribution of the training samples $\tilde{p}_e$ cannot correspond in this case with a feasible distribution in $\Delta(\mathcal{Z})$, because $T(q) = \tilde{T}(\tilde{p}_e)$ with $\tilde{T} = I$ requires that q takes both positive and negative values. Specifically, if $\mathcal{Z} = \{-1, +1\}$,

$$T = I \otimes \begin{bmatrix} 1-\rho^- & \rho^+ \\ \rho^- & 1-\rho^+ \end{bmatrix} \tag{7}$$

with ρ^- and ρ^+ the probabilities of erroneous labelling in training when the actual label is -1 and $+1$, respectively. The methods presented in [16, 24] transform original loss function $L(\hat{y}, y)$ as $\widetilde{L}(\hat{y}, y)$ with

$$\begin{bmatrix} \widetilde{L}(\hat{y}, -1) \\ \widetilde{L}(\hat{y}, -1) \end{bmatrix} = \begin{bmatrix} 1-\rho^- & \rho^+ \\ \rho^- & 1-\rho^+ \end{bmatrix}^{-1} \begin{bmatrix} L(\hat{y}, -1) \\ L(\hat{y}, -1) \end{bmatrix} = \frac{1}{1-\rho^- - \rho^+} \begin{bmatrix} 1-\rho^+ & -\rho^- \\ -\rho^+ & 1-\rho^- \end{bmatrix} \begin{bmatrix} L(\hat{y}, -1) \\ L(\hat{y}, -1) \end{bmatrix}$$

and then obtain classification rules by minimizing the expected loss with respect to empirical distributions $\tilde{p}_e$. However, if $T(q) = \tilde{p}_e$ and $x^{(i)}$ is an instance incorrectly labelled in training as $\tilde{y} = -1$, then (7) implies that

$$q(x^{(i)}, y = 1) = -\frac{\rho^+}{n(1-\rho^- - \rho^+)} \tag{8}$$

that can be significantly smaller than zero for moderate training sizes. This example illustrates that ERM can be inadequate for noisy labels, since it determines an optimal classification rule with respect to a measure that is not a probability measure.

The presented framework can also enable the development of novel supervision schemes of practical interest. For instance, supervision schemes in which labels at training are more precise than labels at prediction (e.g., multi-option classification with precise training labels) can be seen as examples of the proposed framework with $\mathrm{B} = \mathrm{Z}$ and $\tilde{T} = I \otimes T_{\mathrm{Y}\mid\tilde{\mathrm{Y}}} \in \Delta(\tilde{\mathcal{Z}}, \mathcal{Z})$. Additionally, note that the proposed framework can encompass combinations of the schemes described above. For instance, supervision schemes in which instances at prediction and labels at training are less precise than those at training and prediction, respectively, can be seen as examples of the proposed framework with $\mathrm{B} = (\mathrm{X}, \tilde{\mathrm{Y}})$, $T = I \otimes T_{\tilde{\mathrm{Y}}\mid\mathrm{Y}} \in \Delta(\mathcal{Z}, \mathcal{B})$, and $\tilde{T} = T_{\mathrm{X}\mid\tilde{\mathrm{X}}} \otimes I \in \Delta(\tilde{\mathcal{Z}}, \mathcal{B})$.

Other current techniques such as those developed under the paradigm of "covariate shift" exploit a specific relationship between the probability distributions of examples at prediction and training [22]. Those techniques assume that variables at prediction and training share the same conditional distribution of labels given instances, but instances at prediction and training have different marginal distributions $p(x)$ and $\tilde{p}(x)$, i.e., $\mathcal{X} = \tilde{\mathcal{X}}$, $\mathcal{Y} = \tilde{\mathcal{Y}}$, and

$$p(x, y) = \tilde{p}(x, y) \frac{p(x)}{\tilde{p}(x)} \,. \tag{9}$$

Such techniques use samples of instances at prediction and training to estimate the function $p(x)/\tilde{p}(x)$, and determine the classification rule using a ERM that weights training samples according to the estimated function. Note that (9) can be thought of as a mapping of $\tilde{p}$ to p. However, unlike the proposed probabilistic transformations T and $\widetilde{T}$, such mapping depends on the specific probability distributions followed by prediction and training instances so its usage requires to estimate such mapping for each specific probability distributions.

4 Supervision with heterogeneous training samples

In this section we consider supervision cases in which training data is an ensemble of samples with m different types, i.e., $d = (d_1, d_2, \ldots, d_m)$, and, for $i = 1, 2, \ldots, m$, $d_i = (\tilde{z}_i^{(1)}, \tilde{z}_i^{(2)}, \ldots, \tilde{z}_i^{(n_i)})$ where $\tilde{z}_i^{(j)}$ for $j = 1, 2, \ldots, n_i$ are independent instantiations of $\tilde{\mathrm{Z}}_i$. Several current supervision schemes use the following ensembles of training samples:

- Semi-supervised classification [4, 19]: a subset of training examples miss labels.
- TS missing instances [21]: some training examples miss different instance' components.
- Variable quality data [6, 24]: different subsets of training examples are affected by different noise intensities.

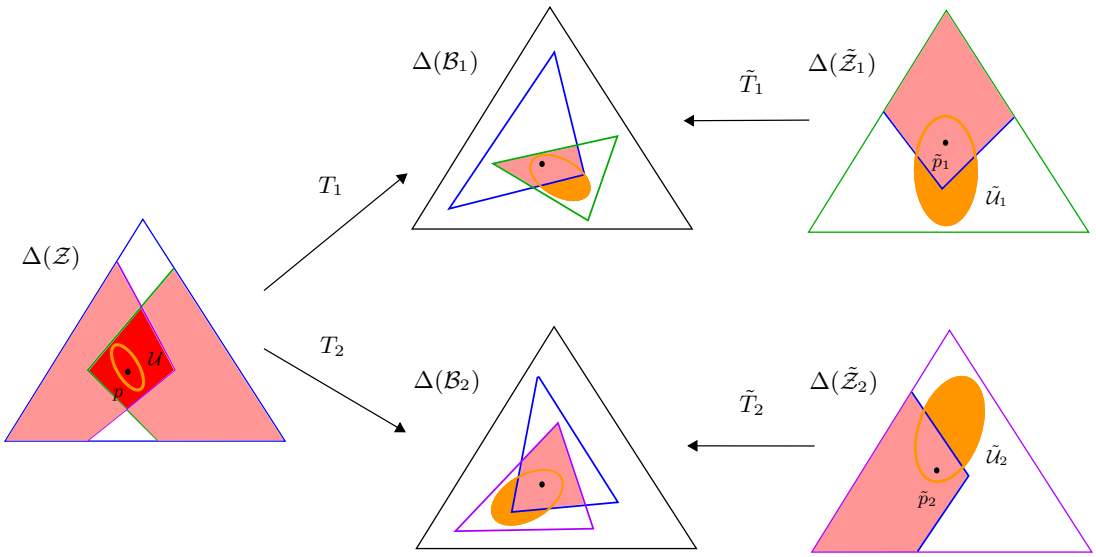

Figure 2. The relationships between prediction and each type of training variables impose structural constraints for feasible distributions (red polygons), and enable to use information from heterogeneous training samples as uncertainty sets (orange ellipses).

- Domain adaptation with multiple sources [15]: different subsets of training examples belong to different but similar domains.

The following shows how heterogenous training samples can be aggregated by further extending the framework presented in previous section. Let, for $i = 1, 2 \ldots, m$, $\mathcal{B}_i$ be a support set, and $T_i \in \Delta(\mathcal{Z}, \mathcal{B}_i)$ and $\tilde{T}_i \in \Delta(\tilde{\mathcal{Z}}_i, \mathcal{B}_i)$ be probabilistic transformations such that $T_i(p) = \tilde{T}_i(\tilde{p}_i)$ for p and $\tilde{p}_i$ the distributions of Z and $\tilde{\mathrm{Z}}_i$, respectively. Analogously to the case described in previous section for only one type of training samples, i.e., $m = 1$, such probabilistic transformations allow to extract the information in heterogeneous and non-standard training data (see Fig. 2). Specifically, feasible distributions and uncertainty sets in $\Delta(\mathcal{Z})$ can be defined as

$$\mathcal{F} = \{q \in \Delta(\mathcal{Z}) : \exists \tilde{q}_i \in \Delta(\tilde{\mathcal{Z}}_i),\ T_i(q) = \tilde{T}_i(\tilde{q}_i) \quad \forall\, i = 1, 2, \ldots, m\}$$

$$\mathcal{U} = \{q \in \Delta(\mathcal{Z}) : \sum_{i=1}^{m} w_i \psi(T_i(q), \tilde{T}_i(\tilde{p}_{e_i})) < \varepsilon\}$$

with $w_i > 0$ a parameter weighting the discrepancy in each $\Delta(\mathcal{B}_i)$, e.g., $w_i \propto \sqrt{n_i}$. Therefore, learning from non-standard heterogeneous training data $d_1, d_2, \ldots, d_m$ can be approached by GRRM generalizing equation (6) as

$$\min_{q \in \mathcal{F}} \sum_{i=1}^{m} w_i \psi\big(T_i(q), \tilde{T}_i(\tilde{p}_{e_i})\big) - \lambda H(q) \tag{10}$$

where $\lambda > 0$ is a regularization parameter.

Table 2 shows how different current supervision schemes with heterogeneous training samples can be addressed under the proposed framework. In semi-supervision and TS missing instances, samples in one subset of the training samples follow the same distribution as those at prediction stage, i.e., $\mathrm{B}_1 = \tilde{\mathrm{Z}}_1 = \mathrm{Z}$, while the remaining training samples are less precise than those at prediction, i.e., $\mathrm{B}_i = \tilde{\mathrm{Z}}_i$ and $T_i = T_{\tilde{\mathrm{Z}}_i|\mathrm{Z}} \in \Delta(\mathcal{Z}, \tilde{\mathcal{Z}}_i)$ for $i > 1$. In particular, for TS missing instances, training samples can be categorized in terms of the instance component that is missing with $\bar{x}_i = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_r)$. In other supervision schemes, such as variable quality data or domain adaptation with multiple sources, the training samples subsets are affected by different label noises ($T_i = I \otimes T_{\tilde{\mathrm{Y}}_i|\mathrm{Y}}$) or belong to different domains with a common representation ($T_i = \tilde{T}_i = T_{\mathrm{B}|\mathrm{Z}_i}$), respectively.

The proposed framework can also enable the development of novel supervision schemes that aggregate general ensembles of training samples, such as those described in fourth case study in Section 5. These new supervision schemes could be specially suitable for environments of open collaboration where each participant in the annotation process could choose a type of contribution based on resources, commitment, remuneration, etc. For instance, different groups of participants could choose to use high- or low-resolution instances, to annotate examples quickly or meticulously, etc.

5 Experiments

This section shows the feasibility of the general framework proposed to encompass multiple existing schemes as well as to enable novel types of supervision. Specifically, we consider four experimentation case studies: two well-studied non-standard supervision schemes, and two novel non-standard supervision schemes. We solved the convex optimization problems (6) and (10) using CVX package [10] with entropy given by (4). As in [9], the distributions considered have instances support that coincides with that of the empirical distribution, and we use the discrepancy given by (3).

Table 4 shows the estimated accuracy of proposed GRRM for two existing supervision schemes (noisy labels and semi-supervision) in comparison with several representative methods using the 3 UCI datasets that are used in both [14] and [19] (see details in Table 3). In these two case studies we used (3) with statistic

$$t(z) = (\theta_-(y), \theta_-(y)x, \theta_+(y), \theta_+(y)x)$$

where $\theta = (\theta_-, \theta_+)$ is the one-hot encoding [9] of the class y and the step (1) is solved by a support vector machine (SVM) with weights given by the solutions of (6) and (10). For noisy labels we compare

Table 2. Current heterogeneous supervision schemes

Supervision scheme	Training samples types $\tilde{\mathrm{Z}}_i$	Bridges B_i	Prob. transformations
Semi-supervision	$\tilde{\mathrm{Z}}_1 = \mathrm{Z} = (\mathrm{X}, \mathrm{Y})$ $\tilde{\mathrm{Z}}_2 = \mathrm{X}$	$\mathrm{B}_1 = \mathrm{Z} = \tilde{\mathrm{Z}}_1$ $\mathrm{B}_2 = \tilde{\mathrm{Z}}_2$	$T_1 = \tilde{T}_1 = I$ $T_2 = T_{\mathrm{X}\vert\mathrm{Z}}, \tilde{T}_2 = I$
TS missing instances	$\tilde{\mathrm{Z}}_1 = \mathrm{Z} = (\mathrm{X}, \mathrm{Y})$ $\tilde{\mathrm{Z}}_{i+1} = (\bar{\mathrm{X}}_i, \mathrm{Y})$	$\mathrm{B}_1 = \mathrm{Z} = \tilde{\mathrm{Z}}_1$ $\mathrm{B}_{i+1} = \tilde{\mathrm{Z}}_{i+1}$	$T_1 = \tilde{T}_1 = I$ $T_{i+1} = T_{\bar{\mathrm{X}}_i\vert\mathrm{X}} \otimes I$ $\tilde{T}_{i+1} = I$
Variable quality data	$\tilde{Z}_i = (\mathrm{X}, \mathrm{Y}_i)$, y_i noisy $\mathrm{Y}_i \neq \mathrm{Y}_j, i \neq j$	$\mathrm{B}_i = \tilde{\mathrm{X}}_i$	$T_i = I \otimes T_{\tilde{\mathrm{Y}}_i\vert\mathrm{Y}}$ $\tilde{T}_i = I$
Domain adaptation with multiple sources	$\tilde{\mathcal{X}}_i = \mathcal{X}, \tilde{\mathrm{X}}_i \neq \mathrm{X}$	General	$T_i = \tilde{T}_i = T_{\mathrm{B}\vert\mathrm{Z}}$

Table 3. Data sets

Name	dim.	instances	num. samples
Diabetes	8	768	(268+,500-)
German	20	1000	(300+,700-)
Heart	13	270	(120+,150-)
Tic-tac-toe	9	958	(626+,332-)

the accuracy of GRRM with that of 4 methods, as reported in [14] (case $\rho^- = 0.1$ and $\rho^+ = 0.3$). For semi-supervision we compare the accuracy of GRRM with that of 3 methods, as reported in [19], as well as method SMIR[4] proposed in [17] (we used 5% and 30% labeled and unlabeled samples, resp.). The results in Table 4 show that GRRM can obtain state-of-the-art accuracies in well-studied non-standard supervision schemes.

Table 4. Accuracy of proposed GRRM for existing supervision schemes.

	Technique	Data set German	Heart	Diabetes
Noisy labels	GRRM	72.6%	78.3%	73.2%
	IW	69.6%	72.1%	71.5%
	LD	70.8%	72.2%	73.2%
	eIW	68.8%	70.1%	74.3%
	StPMKL	67.2%	54.7%	66.5%
Semi-supervision	GRRM	70.0%	77.8%	70.0%
	Lap-TSVM	63.5%	75.8%	63.4%
	Lap-SVM	64.6%	74.3%	63.0%
	TSVM	61.2%	73.7%	60.0%
	SMIR	70.0%	75.1%	68.6%

Fig. 3 and Fig. 4 show the accuracy of proposed GRRM in novel non-standard supervision schemes using the UCI tic-tac-toe dataset. In particular, the board configurations in the 2x2 upper-left block are used as instances to predict the game end, and classification is done by computing labels' conditional probabilities.

[4] Implemented using code in https://github.com/wittawatj/smir

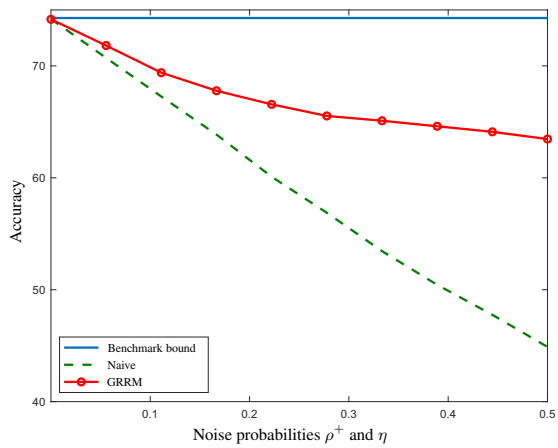

Figure 3. Supervision with noisy labels (training) and noisy instances (prediction).

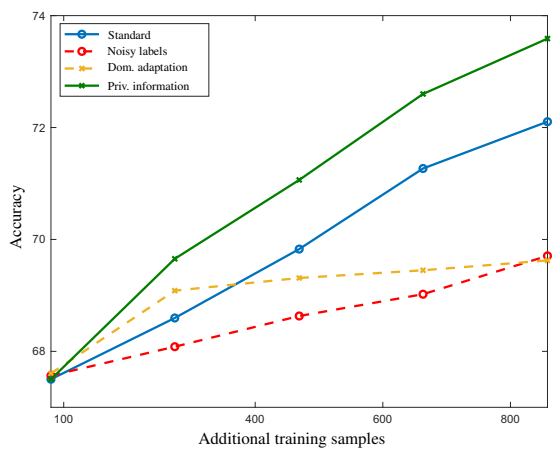

Figure 4. Learning curves using training samples with 4 different types.

The first novel supervision scheme considers noisy labels at training and noisy instances at prediction. We compare classification accuracy with varying probabilities of errors for 3 implementations: benchmark bound obtained by using ERM with noiseless instances and labels, naive ERM that does not account for the noises, and proposed GRRM using (3) with indicator functions of each board case as statistics. The probabilities of incorrectly labeling a "win for x" and a "not win for x" are ρ^+ and ρ^-, respectively, while the probability of an error in reading each board's cell is η. Fig. 3 compares the accuracies obtained varying ρ^+ and η from 0 to 0.5 with $\rho^+ = \eta$ and $\rho^- = \rho^+/2$. It can be observed that proposed GRRM can enable the usage of both noisy labels at training and noisy instances at prediction even when they are severely affected by noise.

The second novel supervision scheme aggregates training samples with 4 different types: standard supervision, noisy labels ($\rho^- = 0.1$, $\rho^+ = 0.3$), domain adaptation with the middle vertical 3x1 block as instances, and privileged information with all cells except the upright and low-left corners as instances. Fig. 4 compares the accuracies obtained by proposed GRRM using different amounts of training samples for each type. The leftmost points in the curves show the accuracy obtained aggregating 80 samples of each type, and the remaining points show how accuracy increases by increasing the number of training samples of different types while keeping the others fixed. It can be observed that the proposed GRRM can aggregate training samples with different types. As expected, the accuracy increases faster by adding more informative training samples (standard and privileged information) than by adding less informative training samples (noisy labels and domain adaptation). These heterogenous supervision schemes can improve the accuracy vs cost trade-off in training stages by enabling the aggregation of multiple samples' types with different acquisition costs and information contents.

6 Conclusion

The paper presents a unifying framework and learning techniques for supervised classification with non-standard and heterogenous training data. The introduced methodology of generalized robust risk minimization (GRRM) can enable to develop learning algorithms for current and novel supervision schemes in a unified manner. The results presented can lead to new learning scenarios able to balance cost vs accuracy trade-offs of training stages, and seamlessly aggregate ensembles of training samples with assorted types and qualities.

Acknowledgements

We thank the Spanish Ministry of Science through Ramon y Cajal under Grant RYC-2016-19383 and TIN2017-82626-R, the BCAM's Severo Ochoa Excellence Accreditation SEV-2017-0718, the Basque Government through the ELKARTEK and BERC 2018-2021 programmes, and the 2018 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation.

REFERENCES

[1] Charalambos D. Aliprantis and Kim C. Border, *Infinite dimensional analysis*, Springer-Verlag, Berlin, 1994.

[2] Kaiser Asif, Wei Xing, Sima Behpour, and Brian D. Ziebart, 'Adversarial cost-sensitive classification', in *Conference on Uncertainty in Artificial Intelligence*, pp. 92–101, (2015).

[3] Dimitris Bertsimas and Martin S. Copenhaver, 'Characterization of the equivalence of robustification and regularization in linear and matrix regression', *European Journal of Operational Research*, **270**(3), 931–942, (2017).

[4] Olivier Chapelle and Alexander Zien, 'Semi-supervised classification by low density separation', in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 57–64, (2005).

[5] Minmin Chen, Kilian Q. Weinberger, and John C. Blitzer, 'Co-training for domain adaptation', in *Advances in Neural Information Processing Systems*, pp. 2456–2464, (2011).

[6] Koby Crammer, Michael Kearns, and Jennifer Wortman, 'Learning from data of variable quality', in *Advances in Neural Information Processing Systems*, pp. 219–226, (2006).

[7] Ofer Dekel, Ohad Shamir, and Lin Xiao, 'Learning to classify with missing and corrupted features', *Machine Learning*, **81**(2), 149–178, (2010).

[8] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio, 'Regularization networks and support vector machines', *Advances in computational mathematics*, **13**(1), 1–50, (2000).

[9] Farzan Farnia and David Tse, 'A minimax approach to supervised learning', in *Advances in Neural Information Processing Systems*, pp. 4240–4248, (2016).

[10] Michael Grant, Stephen Boyd, and Yinyu Ye, 'Disciplined convex programming', in *Global Optimization: From Theory to Implementation*, eds., L. Liberti and N. Maculan, Nonconvex Optimization and its Applications, 155–210, Springer, (2006).

[11] Peter D. Grünwald and A. Philip Dawid, 'Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory', *Annals of Statistics*, **32**(4), 1367–1433, (2004).

[12] Rong Jin and Zoubin Ghahramani, 'Learning with multiple labels', in *Advances in Neural Information Processing Systems*, pp. 921–928, (2003).

[13] Jaeho Lee and Maxim Raginsky, 'Minimax statistical learning with Wasserstein distances', in *Advances in Neural Information Processing Systems*, pp. 2692–2701, (2018).

[14] Tongliang Liu and Dacheng Tao, 'Classification with noisy labels by importance reweighting', *IEEE Transactions on pattern analysis and machine intelligence*, **38**(3), 447–461, (2016).

[15] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh, 'Domain adaptation with multiple sources', in *Advances in Neural Information Processing Systems*, pp. 1041–1048, (2009).

[16] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari, 'Learning with noisy labels', in *Advances in Neural Information Processing Systems*, pp. 1196–1204, (2013).

[17] Gang Niu, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, and Masashi Sugiyama, 'Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning', in *International Conference on Machine Learning*, pp. 10–18, (2013).

[18] Dmitry Pechyony and Vladimir Vapnik, 'On the theory of learnining with privileged information', in *Advances in Neural Information Processing Systems*, pp. 1894–1902, (2010).

[19] Zhiquan Qi, Yingjie Tian, and Yong Shi, 'Laplacian twin support vector machine for semi-supervised classification', *Neural Networks*, **35**, 46–53, (2012).

[20] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani, 'Regularization via mass transportation', *arXiv preprint*, **arXiv:1710.10016**, (2017).

[21] Pannagadatta K. Shivaswamy, Chiranjib Bhattacharyya, and Alexander J. Smola, 'Second order cone programming approaches for handling missing and uncertain data', *Journal of Machine Learning Research*, **7**, 1283–1314, (2006).

[22] Masashi Sugiyama and Motoaki Kawanabe, *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*, MIT press, Cambridge, US, 2012.

[23] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou, 'Multi-label learning with weak label', in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pp. 593–598, (2010).

[24] Brendan van Rooyen and Robert C. Williamson, 'A theory of learning with corrupted labels', *Journal of Machine Learning Research*, **18**, 1–50, (July 2018).

[25] Vapnik Vladimir, *Statistical learning theory*, Wiley, New York, 1998.