Journal of Machine Learning Research 16 (2015) 2023-2049 Submitted 7/15; Published 9/15

In memory of Alexey Chervonenkis

Learning Using Privileged Information: Similarity Control and Knowledge Transfer

Vladimir Vapnik VLADIMIR.VAPNIK@GMAIL.COM
Columbia University
New York, NY 10027, USA
Facebook AI Research
New York, NY 10017, USA

Rauf Izmailov RIZMAILOV@APPCOMSCI.COM
Applied Communication Sciences
Basking Ridge, NJ 07920-2021, USA

Editor: Alex Gammerman and Vladimir Vovk

Abstract

This paper describes a new paradigm of machine learning, in which Intelligent Teacher is involved. During training stage, Intelligent Teacher provides Student with information that contains, along with classification of each example, additional privileged information (for example, explanation) of this example. The paper describes two mechanisms that can be used for significantly accelerating the speed of Student's learning using privileged information: (1) correction of Student's concepts of similarity between examples, and (2) direct Teacher-Student knowledge transfer.

Keywords: intelligent teacher, privileged information, similarity control, knowledge transfer, knowledge representation, frames, support vector machines, SVM+, classification, learning theory, kernel functions, similarity functions, regression

1. Introduction

During the last fifty years, a strong machine learning theory has been developed. This theory (see Vapnik and Chervonenkis, 1974, Vapnik, 1995, Vapnik, 1998, Chervonenkis, 2013) includes:

- The necessary and sufficient conditions for consistency of learning processes.
- The bounds on the rate of convergence, which, in general, cannot be improved.
- The new inductive principle called Structural Risk Minimization (SRM), which always converges to the best possible approximation in the given set of functions[1].

1. Let a set S of functions $f(x, \alpha), \alpha \in \Lambda$ be given. We introduce a structure $S_1 \subset S_2 \subset ... \subset S$ on this set, where S_k is the subset of functions with VC dimension k. Consider training set $(x_1, y_1), \ldots, (x_\ell, y_\ell)$. In the SRM framework, by choosing an element S_k and a function in this element to minimize the CV bound for samples of size ℓ, one chooses functions $f(x, \alpha_\ell) \in S_k$ such that the sequence $\{f(x, \alpha_\ell\}, \ell \to \infty$,

©2015 Vladimir Vapnik and Rauf Izmailov.

- The effective algorithms, such as Support Vector Machines (SVM), that realize the consistency property of SRM principle[2].

The general learning theory appeared to be completed: it addressed almost all standard questions of the statistical theory of inference. However, as always, the devil is in the detail: it is a common belief that human students require far fewer training examples than any learning machine. Why?

We are trying to answer this question by noting that a human Student has an Intelligent Teacher[3] and that Teacher-Student interactions are based not only on brute force methods of function estimation. In this paper, we show that Teacher-Student interactions can include special learning mechanisms that can significantly accelerate the learning process. In order for a learning machine to use fewer observations, it can use these mechanisms as well.

This paper considers a model of learning with the so-called Intelligent Teacher, who supplies Student with intelligent (privileged) information during training session. This is in contrast to the classical model, where Teacher supplies Student only with outcome y for event x.

Privileged information exists for almost any learning problem and this information can significantly accelerate the learning process.

2. Learning with Intelligent Teacher: Privileged Information

The existing machine learning paradigm considers a simple scheme: given a set of training examples, find, in a given set of functions, the one that approximates the unknown decision rule in the best possible way. In such a paradigm, Teacher does not play an important role.

In human learning, however, the role of Teacher is important: along with examples, Teacher provides students with explanations, comments, comparisons, metaphors, and so on. In the paper, we include elements of human learning into classical machine learning paradigm. We consider a learning paradigm called *Learning Using Privileged Information (LUPI)*, where, at the training stage, Teacher provides additional information x^* about training example x.

The crucial point in this paradigm is that the privileged information is available only at the training stage (when Teacher interacts with Student) and is not available at the test stage (when Student operates without supervision of Teacher).

In this paper, we consider two mechanisms of Teacher–Student interactions in the framework of the LUPI paradigm:

1. *The mechanism to control Student's concept of similarity between training examples.*

---

strongly uniformly converges to the function $f(x, \alpha_0)$ that minimizes the error rate on the closure of $\cup_{k=1}^{\infty} S_k$ (Vapnik and Chervonenkis, 1974), (Vapnik, 1982), (Devroye et al., 1996), (Vapnik, 1998).

2. Solutions of SVM belong to Reproducing Kernel Hilbert Space (RKHS). Any subset of functions in RKHS with bounded norm has a finite VC dimension. Therefore, SRM with respect to the value of norm of functions satisfies the general SRM model of strong uniform convergence. In SVM, the element of SRM structure is defined by parameter C of SVM algorithm.
3. This is how a Japanese proverb assesses teacher's influence: "Better than a thousand days of diligent study is one day with a great teacher."

2. *The mechanism to transfer knowledge from the space of privileged information (space of Teacher's explanations) to the space where decision rule is constructed.*

The first mechanism (Vapnik, 2006) was introduced in 2006 using SVM+ method. Here we reinforce SVM+ by constructing a parametric family of methods $\text{SVM}_\Delta+$; for $\Delta = \infty$, the method $\text{SVM}_\Delta+$ is equivalent to SVM+. The first experiments with privileged information using SVM+ method were described in Vapnik and Vashist (2009); later, the method was applied to a number of other examples (Sharmanska et al., 2013; Ribeiro et al., 2012; Liang and Cherkassky, 2008).

The second mechanism was introduced recently (Vapnik and Izmailov, 2015b).

2.1 Classical Model of Learning

Formally, the classical paradigm of machine learning is described as follows: given a set of iid pairs (training data)

$$(x_1, y_1), ..., (x_\ell, y_\ell), \quad x_i \in X, \quad y_i \in \{-1, +1\}, \tag{1}$$

generated according to a fixed but unknown probability measure $P(x, y)$, find, in a given set of indicator functions $f(x, \alpha), \alpha \in \Lambda$, the function $y = f(x, \alpha_*)$ that minimizes the probability of incorrect classifications (incorrect values of $y \in \{-1, +1\}$). In this model, each vector $x_i \in X$ is a description of an example generated by Nature according to an unknown generator $P(x)$ of random vectors x_i, and $y_i \in \{-1, +1\}$ is its classification defined according to a conditional probability $P(y|x)$. The goal of Learning Machine is to find the function $y = f(x, \alpha_*)$ that guarantees the smallest probability of incorrect classifications. That is, the goal is to find the function which minimizes the risk functional

$$R(\alpha) = \frac{1}{2} \int |y - f(x, \alpha)| dP(x, y) \tag{2}$$

in the given set of indicator functions $f(x, \alpha),\ \alpha \in \Lambda$ when the probability measure $P(x, y) = P(y|x)P(x)$ is unknown but training data (1) are given.

2.2 LUPI Paradigm of Learning

The LUPI paradigm describes a more complex model: given a set of iid triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell), \quad x_i \in X, \quad x_i^* \in X^*, \quad y_i \in \{-1, +1\}, \tag{3}$$

generated according to a fixed but unknown probability measure $P(x, x^*, y)$, find, in a given set of indicator functions $f(x, \alpha), \alpha \in \Lambda$, the function $y = f(x, \alpha_*)$ that guarantees the smallest probability of incorrect classifications (2).

In the LUPI paradigm, we have exactly the same goal of minimizing (2) as in the classical paradigm, i.e., to find the best classification function in the admissible set. However, during the training stage, we have more information, i.e., we have triplets (x, x^*, y) instead of pairs (x, y) as in the classical paradigm. The additional information $x^* \in X^*$ belongs to space X^*, which is, generally speaking, different from X. For any element (x_i, y_i) of training example generated by Nature, Intelligent Teacher generates the privileged information x_i^* using some (unknown) conditional probability function $P(x_i^*|x_i)$.

In this paper, we first illustrate the work of these mechanisms on SVM algorithms; after that, we describe their general nature.

Since the additional information is available only for the training set and *is not* available for the test set, it is called *privileged information* and the new machine learning paradigm is called *Learning Using Privileged Information*.

Next, we consider three examples of privileged information that could be generated by Intelligent Teacher.

Example 1. Suppose that our goal is to find a rule that predicts the outcome y of a surgery in three weeks after it, based on information x available before the surgery. In order to find the rule in the classical paradigm, we use pairs (x_i, y_i) from previous patients.

However, for previous patients, there is also additional information x^* about procedures and complications during surgery, development of symptoms in one or two weeks after surgery, and so on. Although this information is not available *before* surgery, it does exist in historical data and thus can be used as privileged information in order to construct a rule that is better than the one obtained without using that information. The issue is how large an improvement can be achieved.

Example 2. Let our goal be to find a rule $y = f(x)$ to classify biopsy images x into two categories y: cancer ($y = +1$) and non-cancer ($y = -1$). Here images are in a pixel space X, and the classification rule has to be in the same space. However, the standard diagnostic procedure also includes a pathologist's report x^* that describes his/her impression about the image in a high-level holistic language X^* (for example, "aggressive proliferation of cells of type A among cells of type B" etc.).

The problem is to use the pathologist's reports x^* as privileged information (along with images x) in order to make a better classification rule for images x just in pixel space X. (Classification by a pathologist is a time-consuming procedure, so fast decisions during surgery should be made without consulting him or her).

Example 3. Let our goal be to predict the direction of the exchange rate of a currency at the moment t. In this problem, we have observations about the exchange rates before t, and we would like to predict if the rate will go up or down at the moment $t + \Delta$. However, in the historical market data we also have observations about exchange rates *after* moment t. Can this future-in-the-past privileged information be used for construction of a better prediction rule?

To summarize, privileged information is ubiquitous: it usually exists for almost any machine learning problem.

Section 4 describes the first mechanism that allows one to take advantage of privileged information by controlling Student's concepts of similarity between training examples. Section 5 describes examples where LUPI model uses similarity control mechanism. Section 6 is devoted to mechanism of knowledge transfer from space of privileged information X^* into decision space X.

However, first in the next Section we describe statistical properties of machine learning that enable the use of privileged information.

3. Statistical Analysis of the Rate of Convergence

According to the bounds developed in the VC theory (Vapnik and Chervonenkis, 1974), (Vapnik, 1998), the rate of convergence depends on two factors: how well the classification rule separates the training data

$$(x_1, y_1), ..., (x_\ell, y_\ell), \quad x \in R^n, \ y \in \{-1, +1\}, \tag{4}$$

and the VC dimension of the set of functions in which the rule is selected.

The theory has two distinct cases:

1. **Separable case:** there exists a function $f(x, \alpha_\ell)$ in the set of functions $f(x, \alpha), \alpha \in \Lambda$ with finite VC dimension h that separates the training data (4) without errors:

   $$y_i f(x_i, \alpha_\ell) > 0 \ \ \forall i = 1, ..., \ell.$$

   In this case, for the function $f(x, \alpha_\ell)$ that minimizes (down to zero) the empirical risk (on training set (4)), the bound

   $$P(yf(x, \alpha_\ell) \leq 0) < O^* \left(\frac{h - \ln \eta}{\ell} \right)$$

   holds true with probability $1 - \eta$, where $P(yf(x, \alpha_\ell) \leq 0)$ is the probability of error for the function $f(x, \alpha_\ell)$ and h is the VC dimension of the admissible set of functions. Here O^* denotes order of magnitude up to logarithmic factor.

2. **Non-separable case:** there is no function in $f(x, \alpha), \ \alpha \in \Lambda$ finite VC dimension h that can separate data (4) without errors. Let $f(x, \alpha_\ell)$ be a function that minimizes the number of errors on (4). Let $\nu(\alpha_\ell)$ be its error rate on training data (4). Then, according to the VC theory, the following bound holds true with probability $1 - \eta$:

   $$P(yf(x, \alpha_\ell) \leq 0) < \nu(\alpha_\ell) + O^* \left(\sqrt{\frac{h - \ln \eta}{\ell}} \right).$$

In other words, in the separable case, the rate of convergence has the order of magnitude $1/\ell$; in the non-separable case, the order of magnitude is $1/\sqrt{\ell}$. The difference between these rates[4] is huge: the same order of bounds requires 320 training examples versus 100,000 examples. Why do we have such a large gap?

3.1 Key Observation: SVM with Oracle Teacher

Let us try to understand why convergence rates for SVMs differ so much for separable and non-separable cases. Consider two versions of the SVM method for these cases.

SVM method first maps vectors x of space X into vectors z of space Z and then constructs a separating hyperplane in space Z. If training data can be separated with no error (the so-called separable case), SVM constructs (in space Z that we, for simplicity,

4. The VC theory also gives a more accurate estimate of the rate of convergence; however, the scale of difference remains essentially the same.

consider as an N-dimensional vector space R^N) a maximum margin separating hyperplane. Specifically, in the separable case, SVM minimizes the functional

$$\mathcal{T}(w) = (w, w)$$

subject to the constraints

$$(y_i(w, z_i) + b) \geq 1, \quad \forall i = 1, ..., \ell;$$

whereas in the non-separable case, SVM minimizes the functional

$$\mathcal{T}(w) = (w, w) + C \sum_{i=1}^{\ell} \xi_i$$

subject to the constraints

$$(y_i(w, z_i) + b) \geq 1 - \xi_i, \quad \forall i = 1, ..., \ell,$$

where $\xi_i \geq 0$ are slack variables. That is, in the separable case, SVM uses ℓ observations for estimation of N coordinates of vector w, whereas in the nonseparable case, SVM uses ℓ observations for estimation of $N + \ell$ parameters: N coordinates of vector w and ℓ values of slacks ξ_i. Thus, in the non-separable case, the number $N + \ell$ of parameters to be estimated is always larger than the number ℓ of observations; it does not matter here that most of slacks will be equal to zero: SVM still has to estimate all ℓ of them. Our guess is that the difference between the corresponding convergence rates is due to the number of parameters SVM has to estimate.

To confirm this guess, consider the SVM with *Oracle Teacher* (Oracle SVM). Suppose that Teacher can supply Student with the values of slacks as privileged information: during training session, Student is supplied with triplets

$$(x_1, \xi_1^0, y_1), ..., (x_\ell, \xi_\ell^0, y_\ell),$$

where $\xi_i^0,\ i = 1, ..., \ell$ are the slacks for the Bayesian decision rule. Therefore, in order to construct the desired rule using these triplets, the SVM has to minimize the functional

$$\mathcal{T}(w) = (w, w)$$

subject to the constraints

$$(y_i(w, z_i) + b) \geq r_i, \quad \forall i = 1, ..., \ell,$$

where we have denoted

$$r_i = 1 - \xi_i^0, \quad \forall i = 1, ..., \ell.$$

One can show that the rate of convergence is equal to $O^*(1/\ell)$ for Oracle SVM. The following (slightly more general) proposition holds true (Vapnik and Vashist, 2009).

Proposition 1. *Let $f(x, \alpha_0)$ be a function from the set of indicator functions $f(x, \alpha)$, with $\alpha \in \Lambda$ with VC dimension h that minimizes the frequency of errors (on this set) and let*

$$\xi_i^0 = \max\{0, (1 - f(x_i, \alpha_0))\}, \quad \forall i = 1, ..., \ell.$$

Then the error probability $p(\alpha_\ell)$ for the function $f(x, \alpha_\ell)$ that satisfies the constraints

$$y_i f(x, \alpha) \geq 1 - \xi_i^0, \quad \forall i = 1, ..., \ell$$

is bounded, with probability $1 - \eta$, as follows:

$$p(\alpha_\ell) \leq P(1 - \xi_0 < 0) + O^* \left(\frac{h - \ln \eta}{\ell} \right).$$

3.2 From Ideal Oracle to Real Intelligent Teacher

Of course, real Intelligent Teacher cannot supply slacks: Teacher does not know them. Instead, Intelligent Teacher can do something else, namely:

1. define a space X^* of (correcting) slack functions (it can be different from the space X of decision functions);

2. define a set of real-valued slack functions $f^*(x^*, \alpha^*),\ x^* \in X^*,\ \alpha^* \in \Lambda^*$ with VC dimension h^*, where approximations

$$\xi_i = f^*(x, \alpha^*)$$

of the slack functions[5] are selected;

3. generate privileged information for training examples supplying Student, instead of pairs (4), with triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell). \tag{5}$$

During training session, the algorithm has to simultaneously estimate two functions using triplets (5): the decision function $f(x, \alpha_\ell)$ and the slack function $f^*(x^*, \alpha_\ell^*)$. In other words, the method minimizes the functional

$$\mathcal{T}(\alpha^*) = \sum_{i=1}^{\ell} \max\{0, f^*(x_i^*, \alpha^*)\} \tag{6}$$

subject to the constraints

$$y_i f(x_i, \alpha) > -f^*(x_i^*, \alpha^*), \quad i = 1, ..., \ell. \tag{7}$$

Let $f(x, \alpha_\ell)$ and $f^*(x^*, \alpha_\ell^*)$ be functions that solve this optimization problem. For these functions, the following proposition holds true (Vapnik and Vashist, 2009).

5. Note that slacks ξ_i introduced for the SVM method can be considered as a realization of some function $\xi = \xi(x, \beta_0)$ from a large set of functions (with infinite VC dimension). Therefore, generally speaking, the classical SVM approach can be viewed as estimation of two functions: (1) the decision function, and (2) the slack function, where these functions are selected from two different sets, with finite and infinite VC dimensions, respectively. Here we consider both sets with finite VC dimensions.

Proposition 2. *The solution $f(x, \alpha_\ell)$ of optimization problem (6), (7) satisfies the bounds*

$$P(yf(x, \alpha_\ell) < 0) \leq P(f^*(x^*, \alpha_\ell^*) \geq 0) + O^* \left(\frac{h + h^* - \ln \eta}{\ell} \right)$$

with probability $1 - \eta$, where h and h^* are the VC dimensions of the set of decision functions $f(x, \alpha),\ \alpha \in \Lambda$, and the set of correcting functions $f^*(x^*, \alpha^*),\ \alpha^* \in \Lambda^*$, respectively.

According to Proposition 2, in order to estimate the rate of convergence to the best possible decision rule (in space X) one needs to estimate the rate of convergence of $P\{f^*(x^*, \alpha_\ell^*) \geq 0\}$ to $P\{f^*(x^*, \alpha_0^*) \geq 0\}$ for the best rule $f^*(x^*, \alpha_0^*)$ in space X^*. Note that both the space X^* and the set of functions $f^*(x^*, \alpha_\ell^*), \alpha^* \in \Lambda^*$ are suggested by Intelligent Teacher that tries to choose them in a way that facilitates a fast rate of convergence. The guess is that a really Intelligent Teacher can indeed do that.

As shown in the VC theory, in standard situations, the uniform convergence has the order $O^*(\sqrt{h^*/\ell})$, where h^* is the VC dimension of the admissible set of correcting functions $f^*(x^*, \alpha^*)$, $\alpha* \in \Lambda^*$. However, for special privileged space X^* and corresponding functions $f^*(x^*, \alpha^*), \alpha^* \in \Lambda^*$, the convergence can be faster (as $O^*([1/\ell]^\delta),\ \delta > 1/2$).

A well-selected privileged information space X^* and Teacher's explanation $P(x^*|x)$ along with sets $\{f(x, \alpha_\ell), \alpha \in \Lambda\}$ and $\{f^*(x^*, \alpha^*), \alpha^* \in \Lambda^*\}$ engender a convergence that is faster than the standard one. The skill of Intelligent Teacher is being able to select of the proper space X^*, generator $P(x^*|x)$, set of functions $f(x, \alpha_\ell), \alpha \in \Lambda$, and set of functions $f^*(x^*, \alpha^*), \alpha^* \in \Lambda^*$: that is what differentiates good teachers from poor ones.

4. Similarity Control in LUPI Paradigm

4.1 $\text{SVM}_\Delta+$ for Similarity Control in LUPI Paradigm

In this section, we extend SVM method of function estimation to the method called SVM+, which allows one to solve machine learning problems in the LUPI paradigm (Vapnik, 2006). The $\text{SVM}_\varepsilon+$ method presented below is a reinforced version of the one described in Vapnik (2006) and used in Vapnik and Vashist (2009).

Consider the model of learning with Intelligent Teacher: given triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell),$$

find in the given set of functions the one that minimizes the probability of incorrect classifications in space X.

As in standard SVM, we map vectors $x_i \in X$ onto the elements z_i of the Hilbert space Z, and map vectors x_i^* onto elements z_i^* of another Hilbert space Z^* obtaining triples

$$(z_1, z_1^*, y_1), ..., (z_\ell, z_\ell^*, y_\ell).$$

Let the inner product in space Z be (z_i, z_j), and the inner product in space Z^* be (z_i^*, z_j^*).

Consider the set of decision functions in the form

$$f(x) = (w, z) + b,$$

where w is an element in Z, and consider the set of correcting functions in the form

$$\xi^*(x^*, y) = [y((w^*, z^*) + b^*)]_+,$$

where w^* is an element in Z^* and $[u]_+ = \max\{0, u\}$.

Our goal is to we minimize the functional

$$\mathcal{T}(w, w^*, b, b^*) = \frac{1}{2}[(w, w) + \gamma(w^*, w^*)] + C\sum_{i=1}^{\ell}[y_i((w^*, z_i^*) + b^*)]_+$$

subject to the constraints

$$y_i[(w, z_i) + b] \geq 1 - [y_i((w^*, z_i^*) - b^*)]_+.$$

The structure of this problem mirrors the structure of the primal problem for standard SVM. However, due to the elements $[u_i]_+ = \max\{0, u_i\}$ that define both the objective function and the constraints here we faced non-linear optimization problem.

To find the solution of this optimization problem, we approximate this non-linear optimization problem with the following quadratic optimization problem: minimize the functional

$$\mathcal{T}(w, w^*, b, b^*) = \frac{1}{2}[(w, w) + \gamma(w^*, w^*)] + C\sum_{i=1}^{\ell}[y_i((w^*, z_i^*) + b^*) + \zeta_i] + \Delta C\sum_{i=1}^{\ell}\zeta_i \qquad (8)$$

(here $\Delta > 0$ is the parameter of approximation[6]) subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - y_i((w^*, z^*) + b^*) - \zeta_i, \quad i = 1, ..., \ell, \qquad (9)$$

the constraints

$$y_i((w^*, z_i^*) + b^*) + \zeta_i \geq 0, \quad \forall i = 1, ..., \ell, \qquad (10)$$

and the constraints

$$\zeta_i \geq 0, \quad \forall i = 1, ..., \ell. \qquad (11)$$

To minimize the functional (8) subject to the constraints (10), (11), we construct the Lagrangian

$$\mathcal{L}(w, b, w^*, b^*, \alpha, \beta) = \qquad (12)$$

$$\frac{1}{2}[(w, w) + \gamma(w^*, w^*)] + C\sum_{i=1}^{\ell}[y_i((w^*, z_i^*) + b^*) + (1 + \Delta)\zeta_i] - \sum_{i=1}^{\ell}\nu_i\zeta_i -$$

$$\sum_{i=1}^{\ell}\alpha_i\left[y_i[(w, z_i) + b] - 1 + [y_i((w^*, z_i^*) + b^*) + \zeta_i]\right] - \sum_{i=1}^{\ell}\beta_i[y_i((w^*, z_i^*) + b^*) + \zeta_i],$$

where $\alpha_i \geq 0,\ \beta_i \geq 0,\ \nu_i \geq 0,\ i = 1, ..., \ell$ are Lagrange multipliers.

To find the solution of our quadratic optimization problem, we have to find the saddle point of the Lagrangian (the minimum with respect to w, w^*, b, b^* and the maximum with respect to $\alpha_i, \beta_i, \nu_i,\ i = 1, ..., \ell$).

6. In Vapnik (2006), parameter Δ was set at a sufficiently large value.

The necessary conditions for minimum of (12) are

$$\frac{\partial \mathcal{L}(w, b, w^*, b^*, \alpha, \beta)}{\partial w} = 0 \implies w = \sum_{i=1}^{\ell} \alpha_i y_i z_i \tag{13}$$

$$\frac{\partial \mathcal{L}(w, b, w^*, b^*, \alpha, \beta)}{\partial w^*} = 0 \implies w^* = \frac{1}{\gamma} \sum_{i=1}^{\ell} y_i (\alpha_i + \beta_i - C) z_i^* \tag{14}$$

$$\frac{\partial \mathcal{L}(w, b, w^*, b^*, \alpha, \beta)}{\partial b} = 0 \implies \sum_{i=1}^{\ell} \alpha_i y_i = 0 \tag{15}$$

$$\frac{\partial \mathcal{L}(w, b, w^*, b^*, \alpha, \beta)}{\partial b^*} = 0 \implies \sum_{i=1}^{\ell} y_i (C - \alpha_i - \beta_i) = 0 \tag{16}$$

$$\frac{\partial \mathcal{L}(w, b, w^*, b^*, \alpha, \beta)}{\partial \zeta_i} = 0 \implies \alpha_i + \beta_i + \nu_i = (C + \Delta C) \tag{17}$$

Substituting the expressions (13) in (12) and, taking into account (14), (15), (16), and denoting $\delta_i = C - \beta_i$, we obtain the functional

$$\mathcal{L}(\alpha, \delta) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} (z_i, z_j) y_i y_j \alpha_i \alpha_j - \frac{1}{2\gamma} \sum_{i,j=1}^{\ell} (\delta_i - \alpha_i)(\delta_j - \alpha_j)(z_i^*, z_j^*) y_i y_j.$$

To find its saddle point, we have to maximize it subject to the constraints[7]

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \tag{18}$$

$$\sum_{i=1}^{\ell} y_i \delta_i = 0 \tag{19}$$

$$0 \le \delta_i \le C, \quad i = 1, ..., \ell \tag{20}$$

$$0 \le \alpha_i \le \delta_i + \Delta C, \quad i = 1, ..., \ell \tag{21}$$

Let vectors α^0, δ^0 be a solution of this optimization problem. Then, according to (13) and (14), one can find the approximations to the desired decision function

$$f(x) = (w_0, z_i) + b = \sum_{i=1}^{\ell} \alpha_i^* y_i (z_i, z) + b$$

and to the slack function

$$\xi^*(x^*, y) = y_i((w_0^*, z_i^*) + b^*) + \zeta = \sum_{i=1}^{\ell} y_i (\alpha_i^0 - \delta_i^0)(z_i^*, z^*) + b^* + \zeta.$$

7. In SVM+, instead of constraints (21), the constraints $\alpha_i \ge 0$ were used.

The Karush-Kuhn-Tacker conditions for this problem are

$$\begin{cases} \alpha_i^0[y_i[(w_0, z_i) + b + (w_0^*, z_i^*) + b^*] + \zeta_i - 1] = 0 \\ (C - \delta_i^0)[(w_0^*, z_i^*) + b^* + \zeta_i] = 0 \\ \nu_i^0 \zeta_i = 0 \end{cases}$$

Using these conditions, one obtains the value of constant b as

$$b = 1 - y_k(w^0, z_k) = 1 - y_k \left[\sum_{i=1}^{\ell} \alpha_i^0 (z_i, z_k) \right],$$

where (z_k, z_k^*, y_k) is a triplet for which $\alpha_k^0 \neq 0$, $\delta_k^0 \neq C$, $z_i \neq 0$.

As in standard SVM, we use the inner product (z_i, z_j) in space Z in the form of Mercer kernel $K(x_i, x_j)$ and inner product (z_i^*, z_j^*) in space Z^* in the form of Mercer kernel $K^*(x_i^*, x_j^*)$. Using these notations, we can rewrite the SVM$_\Delta$+ method as follows: the decision rule in X space has the form

$$f(x) = \sum_{i=1}^{\ell} y_i \alpha_i^0 K(x_i, x) + b,$$

where $K(\cdot, \cdot)$ is the Mercer kernel that defines the inner product for the image space Z of space X (kernel $K^*(\cdot, \cdot)$ for the image space Z^* of space X^*) and α^0 is a solution of the following dual space quadratic optimization problem: maximize the functional

$$\mathcal{L}(\alpha, \delta) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \frac{1}{2\gamma} \sum_{i,j=1}^{\ell} y_i y_j (\alpha_i - \delta_i)(\alpha_j - \delta_j) K^*(x_i^*, x_j^*)$$

subject to constraints (18) – (21).

Remark. Note that if $\delta_i = \alpha_i$ or $\Delta = 0$, the solution of our optimization problem becomes equivalent to the solution of the standard SVM optimization problem, which maximizes the functional

$$\mathcal{L}(\alpha, \delta) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

subject to constraints (18) – (21) where $\delta_i = \alpha_i$.

Therefore, the difference between SVM$_\Delta$+ and SVM solutions is defined by the last term in objective function (8). In SVM method, the solution depends only on the values of pairwise similarities between training vectors defined by the Gram matrix K of elements $K(x_i, x_j)$ (which defines similarity between vectors x_i and x_j). The SVM$_\Delta$+ solution is defined by objective function (8) that uses two expressions of similarities between observations: one ($K(x_i, x_j)$ for x_i and x_j) that comes from space X and another one ($K^*(x_i^*, x_j^*)$ for x_i^* and x_j^*) that comes from space of privileged information X^*. That is how Intelligent Teacher changes the optimal solution by correcting the concepts of similarity.

The last term in equation (8) defines the instrument for Intelligent Teacher to control the concept of similarity of Student.

Efficient computational implementation of this SVM+ algorithm for classification and its extension for regression can be found in Pechyony et al. (2010) and Vapnik and Vashist (2009), respectively.

4.1.1 Simplified Approach

The described method $SVM_\Delta+$ requires to minimize the quadratic form $\mathcal{L}(\alpha, \delta)$ subject to constraints (18) – (21). For large ℓ it can be a challenging computational problem. Consider the following approximation. Let

$$f^*(x^*, \alpha^*_\ell) = \sum_{i=1}^{\ell} \alpha^*_i K^*(x^*_i, x) + b^*$$

be be an SVM solution in space X^* and let

$$\xi^*_i = [1 - f^*(x^*, \alpha^*_\ell) - b^*]_+$$

be the corresponding slacks. Let us use the linear function

$$\xi_i = t\xi^*_i + \zeta_i, \quad \zeta_i \geq 0$$

as an approximation of slack function in space X. Now we minimize the functional

$$(w, w) + C \sum_{i=1}^{\ell} (t\xi^*_i + (1 + \Delta)\zeta_i), \quad \Delta \geq 0$$

subject to the constraints

$$y_i((w, z_i) + b) > 1 - t\xi^*_i + \zeta_i,$$
$$t > 0, \quad \zeta_i \geq 0, \ i = 1, ..., \ell$$

(here z_i is Mercer mapping of vectors x_i in RKHS).

The solution of this quadratic optimization problem defines the function

$$f(x, \alpha_\ell) = \sum_{i=1}^{\ell} \alpha_i K(x_i, x) + b,$$

where α is solution of the following dual problem: maximize the functional

$$R(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0$$

$$\sum_{i=1}^{\ell} \alpha_i \xi^*_i \leq C \sum_{i=1}^{\ell} \xi^*_i$$

$$0 \leq \alpha_i \leq (1 + \Delta)C, \quad i = 1, ..., \ell$$

4.2 General Form of Similarity Control in LUPI Paradigm

Consider the following two sets of functions: the set $f(x,\alpha), \alpha \in \Lambda$ defined in space X and the set $f^*(x^*,\alpha^*), \alpha^* \in \Lambda^*$, defined in space X^*. Let a non-negative convex functional $\Omega(f) \geq 0$ be defined on the set of functions $f(x,\alpha), \alpha \in \Lambda$, while a non-negative convex functional $\Omega^*(f^*) \geq 0$ be defined on the set of functions $f(x^*,\alpha^*), \alpha^* \in \Lambda^*$. Let the sets of functions $\theta(f(x,\alpha)), \alpha \in \Lambda$, and $\theta(f(x^*,\alpha^*)), \alpha^* \in \Lambda^*$, which satisfy the corresponding bounded functionals

$$\Omega(f) \leq C_k$$

$$\Omega^*(f^*) \leq C_k,$$

have finite VC dimensions h_k and h_k, respectively. Consider the structures

$$S_1 \subset ... \subset S_m ....$$

$$S_1^* \subset ... \subset S_m^* ...$$

defined on corresponding sets of functions.

Let iid observations of triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell)$$

be given. Our goal is to find the function $f(x,\alpha_\ell)$ that minimizes the probability of the test error.

To solve this problem, we minimize the functional

$$\sum_{i=1}^{\ell} f^*(x_i^*, \alpha)$$

subject to constraints

$$y_i[f(x,\alpha) + f(x^*,\alpha^*)] > 1$$

and the constraint

$$\Omega(f) + \gamma\Omega(f^*) \leq C_m$$

(we assume that our sets of functions are such that solutions exist).

Then, for any fixed sets S_k and S_k^*, the VC bounds hold true, and minimization of these bounds with respect to both sets S_k and S_k^* of functions and the functions $f(x,\alpha_\ell)$ and $f^*(x^(,\alpha_\ell^*)$ in these sets is a realization of universally consistent SRM principle.

The sets of functions defined in previous section by the Reproducing Kernel Hilbert Space satisfy this model since any subset of functions from RKHS with bounded norm has finite VC dimension according to the theorem about VC dimension of linear bounded functions in Hilbert space[8].

8. This theorem was proven in mid-1970s (Vapnik and Chervonenkis, 1974) and generalized for Banach spaces in early 2000s (Gurvits, 2001; Vapnik, 1998).

5. Transfer of Knowledge Obtained in Privileged Information Space to Decision Space

In this section, we consider the second important mechanism of Teacher-Student interaction: using privileged information for knowledge transfer from Teacher to Student[9].

Suppose that Intelligent Teacher has some knowledge about the solution of a specific pattern recognition problem and would like to transfer this knowledge to Student. For example, Teacher can reliably recognize cancer in biopsy images (in a pixel space X) and would like to transfer this skill to Student.

Formally, this means that Teacher has some function $y = f_0(x)$ that distinguishes cancer ($f_0(x) = +1$ for cancer and $f_0(x) = -1$ for non-cancer) in the pixel space X. Unfortunately, Teacher does not know this function explicitly (it only exists as a neural net in Teacher's brain), so how can Teacher transfer this construction to Student? Below, we describe a possible mechanism for solving this problem; we call this mechanism *knowledge transfer*.

Suppose that Teacher believes in some theoretical model on which the knowledge of Teacher is based. For cancer model, he or she believes that it is a result of uncontrolled multiplication of the cancer cells (cells of type B) that replace normal cells (cells of type A). Looking at a biopsy image, Teacher tries to generate privileged information that reflects his or her belief in development of such process; Teacher may describe the image as:

Aggressive proliferation of cells of type B into cells of type A.

If there are no signs of cancer activity, Teacher may use the description

Absence of any dynamics in the of standard picture.

In uncertain cases, Teacher may write

There exist small clusters of abnormal cells of unclear origin.

In other words, Teacher has developed a special language that is appropriate for description x_i^* of cancer development based on the model he or she believes in. Using this language, Teacher supplies Student with privileged information x_i^* for the image x_i by generating training triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell). \tag{22}$$

The first two elements of these triplets are descriptions of an image in two languages: in language X (vectors x_i in pixel space), and in language X^* (vectors x_i^* in the space of privileged information), developed for Teacher's understanding of cancer model.

Note that the language of pixel space is universal (it can be used for description of many different visual objects; for example, in the pixel space, one can distinguish between male and female faces), while the language used for describing privileged information is very specific: it reflects just a model of cancer development. This has an important consequence:

9. In machine learning, transfer learning refers to the framework, where experience obtained for solving one problem is used (with proper modifications) for solving another problem, related to the previous one; both problems are assumed to be in the same space, with only some parameters being changed. The knowledge transfer considered here is different: it denotes the transfer of knowledge obtained in one (privileged) space to another (decision) space.

the set of admissible functions in space X has to be rich (has a large VC dimension), while the set of admissible functions in space X^* may be not rich (has a small VC dimension).

One can consider two related pattern recognition problems using triplets (22):

1. The problem of constructing a rule $y = f(x)$ for classification of biopsy in the pixel space X using data
$$(x_1, y_1), ..., (x_\ell, y_\ell). \tag{23}$$

2. The problem of constructing a rule $y = f^*(x^*)$ for classification of biopsy in the space X^* using data
$$(x_1^*, y_1), ..., (x_\ell^*, y_\ell). \tag{24}$$

Suppose that language X^* is so good that it allows to create a rule $y = f_\ell^*(x^*)$ that classifies vectors x^* corresponding to vectors x with the same level of accuracy as the best rule $y = f_\ell(x)$ for classifying data in the pixel space[10].

In the considered example, the VC dimension of the admissible rules in a special space X^* is much smaller than the VC dimension of the admissible rules in the universal space X and, since the number of examples ℓ is the same in both cases, the bounds on the error rate for the rule $y = f_\ell^*(x^*)$ in X^* will be better[11] than those for the rule $y = f_\ell(x)$ in X. Generally speaking, the knowledge transfer approach can be applied if the classification rule $y = f_\ell^*(x^*)$ is more accurate than the classification rule $y = f_\ell(x)$ (the empirical error in privileged space is smaller than the empirical error in the decision space).

The following problem arises: how one can use the knowledge of the rule $y = f_\ell^*(x^*)$ in space X^* to improve the accuracy of the desired rule $y = f_\ell(x)$ in space X?

5.1 Knowledge Representation for SVMs

To answer this question, we formalize the concept of representation of the knowledge about the rule $y = f_\ell^*(x^*)$.

Suppose that we are looking for our rule in Reproducing Kernel Hilbert Space (RKHS) associated with kernel $K^*(x_i^*, x^*)$. According to Representer Theorem (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001), such rule has the form

$$f_\ell^*(x^*) = \sum_{i=1}^{\ell} \gamma_i K^*(x_i^*, x^*) + b, \tag{25}$$

where $\gamma_i,\ i = 1, ..., \ell$ and b are parameters.

Suppose that, using data (24), we found a good rule (25) with coefficients $\gamma_i = \gamma_i^*,\ i = 1, ..., \ell$ and $b = b^*$. This is now the knowledge about our classification problem. Let us formalize the description of this knowledge.

Consider three elements of knowledge representation used in Artificial Intelligence (Brachman and Levesque, 2004):

10. The rule constructed in space X^* cannot be better than the best possible rule in space X, since all information originates in space X.
11. According to VC theory, the guaranteed bound on accuracy of the chosen rule depends only on two factors: the frequency of errors on training set and the VC dimension of admissible set of functions.

1. Fundamental elements of knowledge.
2. Frames (fragments) of the knowledge.
3. Structural connections of the frames (fragments) in the knowledge.

We call the *fundamental elements of the knowledge* a limited number of vectors $u_1^*,...,u_m^*$ from space X^* that can approximate well the main part of rule (25). It could be the support vectors or the smallest number of vectors[12] $u_i \in X^*$:

$$f_\ell^*(x^*) - b = \sum_{i=1}^{\ell} \gamma_i^* K^*(x_i^*, x^*) \approx \sum_{k=1}^{m} \beta_k^* K^*(u_k^*, x^*). \tag{26}$$

Let us call the functions $K^*(u_k^*, x^*),\ k = 1, ..., m$ the *frames* (fragments) of knowledge. Our knowledge

$$f_\ell^*(x^*) = \sum_{k=1}^{m} \beta_k^* K^*(u_k^*, x^*) + b$$

is defined as a linear combination of the frames.

5.1.1 Scheme of Knowledge Transfer Between Spaces

In the described terms, knowledge transfer from X^* into X requires the following:

1. To find the fundamental elements of knowledge $u_1^*, ..., u_m^*$ in space X^*.
2. To find frames (m functions) $K^*(u_1^*, x^*), ..., K^*(u_m^*, x^*)$ in space X^*.
3. To find the functions $\phi_1(x), ..., \phi_m(x)$ in space X such that

$$\phi_k(x_i) \approx K^*(u_k^*, x_i^*) \tag{27}$$

holds true for almost all pairs (x_i, x_i^*) generated by Intelligent Teacher that uses some (unknown) generator $P(x^*, x) = P(x^*|x)P(x)$.

Note that the capacity of the set of functions from which $\phi_k(x)$ are to be chosen can be smaller than that of the capacity of the set of functions from which the classification function $y = f_\ell(x)$ is chosen (function $\phi_k(x)$ approximates just one fragment of knowledge, not the entire knowledge, as function $y = f_\ell^*(x^*)$, which is a linear combination (26) of frames). Also, as we will see in the next section, estimates of all the functions $\phi_1(x), ..., \phi_m(x)$ are done using different pairs as training sets of the same size ℓ. That is, we hope that transfer of m fragments of knowledge from space X^* into space X can be done with higher accuracy than estimating the function $y = f_\ell(x)$ from data (23).

After finding images of frames in space X, the knowledge about the rule obtained in space X^* can be approximated in space X as

$$f_\ell(x) \approx \sum_{k=1}^{m} \delta_k \phi_k(x) + b^*,$$

where coefficients $\delta_k = \gamma_k$ (taken from (25)) if approximations (27) are accurate. Otherwise, coefficients δ_k can be estimated from the training data, as shown in Section 6.3.

12. In machine learning, it is called the reduced number of support vectors (Burges, 1996).

5.1.2 Finding the Smallest Number of Fundamental Elements of Knowledge

Let our functions ϕ belong to RKHS associated with the kernel $K^*(x_i^*, x^*)$, and let our knowledge be defined by an SVM method in space X^* with support vector coefficients α_i. In order to find the smallest number of fundamental elements of knowledge, we have to minimize (over vectors $u_1^*, ..., u_m^*$ and values $\beta_1, ..., \beta_m$) the functional

$$R(u_1^*, ..., u_m^*; \beta_1, ..., \beta_m) = \tag{28}$$

$$\left\| \sum_{i=1}^{\ell} y_i \alpha_i K^*(x_i^*, x^*) - \sum_{s=1}^{m} \beta_s K^*(u_s^*, x^*) \right\|^2_{RKHS} =$$

$$\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K^*(x_i^*, x_j^*) - 2 \sum_{i=1}^{\ell} \sum_{s=1}^{m} y_i \alpha_i \beta_s K^*(x_i^*, u_s^*) + \sum_{s,t=1}^{m} \beta_s \beta_t K^*(u_s^*, u_t^*).$$

The last equality was derived from the following property of the inner product for functions in RKHS (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001):

$$\big(K^*(x_i^*, x^*), K(x_j^*, x^*)\big)_{RKHS} = K^*(x_i^*, x_j^*).$$

5.1.3 Smallest Number of Fundamental Elements of Knowledge for Homogeneous Quadratic Kernel

For general kernel functions $K^*(\cdot, \cdot)$, minimization of (28) is a difficult computational problem. However, for the special homogeneous quadratic kernel

$$K^*(x_i^*, x_j^*) = (x_i^*, x_j^*)^2,$$

this problem has a simple exact solution (Burges, 1996). For this kernel, we have

$$R = \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i^*, x_j^*)^2 - 2 \sum_{i=1}^{\ell} \sum_{s=1}^{m} y_i \alpha_i \beta_s (x_i^*, u_s^*)^2 + \sum_{s,t=1}^{m} \beta_s \beta_t (u_s^*, u_t^*)^2. \tag{29}$$

Let us look for solution in set of orthonormal vectors $u_i^*, ..., u_m^*$ for which we can rewrite (29) as follows

$$\hat{R} = \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i^*, x_j^*)^2 - 2 \sum_{i=1}^{\ell} \sum_{s=1}^{m} y_i \alpha_i \beta_s (x_i^*, u_s^*)^2 + \sum_{s=1}^{m} \beta_s^2 (u_s^*, u_s^*)^2. \tag{30}$$

Taking derivative of $\hat{R}$ with respect to u_k^*, we obtain that the solutions u_k^*, $k = 1, ..., m$ have to satisfy the equations

$$\frac{d\hat{R}}{du_k} = -2\beta_k \sum_{i=1}^{\ell} y_i \alpha_i x_i^* x_i^{*T} u_k^* + 2\beta_k^2 u_k^* = 0.$$

Introducing notation

$$S = \sum_{i=1}^{\ell} y_i \alpha_i x_i^* x_i^{*T}, \tag{31}$$

we conclude that the solutions satisfy the equation

$$Su_k^* = \beta_k u_k^*, \quad k = 1, ..., m.$$

Let us chose from the set $u_1^*, ..., u_m^*$ of eigenvectors of the matrix S the vectors corresponding to the largest in absolute values eigenvalues $\beta_1, \ldots, \beta_m$, which are coefficients of expansion of the classification rule on the frames $(u_k, x^*)^2, \quad k = 1, \ldots, m$.

Using (31), one can rewrite the functional (30) in the form

$$\hat{R} = \mathbf{1}^{\mathbf{T}} S_2 \mathbf{1} - \sum_{k=1}^{m} \beta_k^2, \tag{32}$$

where we have denoted by S_2 the matrix obtained from S with its elements $s_{i,j}$ replaced with $s_{i,j}^2$, and by $\mathbf{1}$ we have denoted the $(\ell \times 1)$-dimensional matrix of ones.

Therefore, in order to find the fundamental elements of knowledge, one has to solve the eigenvalue problem for $(n \times n)$-dimensional matrix S and then select an appropriate number of eigenvectors corresponding to eigenvalues with largest absolute values. One chooses such m eigenvectors for which functional (32) is small. The number m does not exceed n (the dimensionality of matrix S).

5.1.4 Finding Images of Frames in Space X

Let us call the conditional expectation function

$$\phi_k(x) = \int K^*(u_k^*, x^*) p(x^*|x)\, dx^*$$

the image of frame $K^*(u_k^*, x^*)$ in space X. To find m image functions $\phi_k(x)$ of the frames $K(u_k^*, x^*),\ k = 1, ..., m$ in space X, we solve the following m regression estimation problems: find the regression function $\phi_k(x)$ in X, $k = 1, \ldots, m$, using data

$$(x_1, K^*(u_k^*, x_1^*)), ..., (x_\ell, K^*(u_k^*, x_\ell^*)), \quad k = 1, \ldots, m, \tag{33}$$

where pairs (x_i, x_i^*) belong to elements of training triplets (22).

Therefore, using fundamental elements of knowledge $u_1^*, ...u_m^*$ in space X^*, the corresponding frames $K^*(u_1^*, x^*), ..., K^*(u_m^*, x^*)$ in space X^*, and the training data (33), one constructs the transformation of the space X into m-dimensional feature space[13]

$$\phi(x) = (\phi_1(x), ...\phi_m(x)),$$

where k-th coordinate of vector function $\phi(x)$ is defined as $\phi_k = \phi_k(x)$.

5.1.5 Algorithms for Knowledge Transfer

1. Suppose that our regression functions can be estimated accurately: for a sufficiently small $\varepsilon > 0$ the inequalities

$$|\phi_k(x_i) - K^*(u_k^*, x_i^*)| < \varepsilon, \quad \forall k = 1, ..., m \quad \text{and} \quad \forall i = 1, ..., \ell$$

13. One can choose any subset from $(m + n)$-dimensional space $(\phi_1(x), ...\phi_m(x)), x^1, ..., x^n)$.

hold true for almost all pairs (x_i, x_i^*) generated according to $P(x^*|y)$. Then the approximation of our knowledge in space X is

$$f(x) = \sum_{k=1}^{m} \beta_k^* \phi_k(x) + b^*,$$

where $\beta_k^*,\ k = 1, ..., m$ are eigenvalues corresponding to eigenvectors $u_1^*, ..., u_m^*$.

2. If, however, ε is not too small, one can use privileged information to employ both mechanisms of intelligent learning: controlling similarity between training examples and knowledge transfer.

In order to describe this method, we denote by vector ϕ_i the m-dimensional vector with coordinates

$$\phi_i = (\phi_1(x_i), ..., \phi_m(x_i))^T.$$

Consider the following problem of intelligent learning: given training triplets

$$(\phi_1, x_1^*, y_1), ..., (\phi_\ell, x_\ell^*, y_\ell),$$

find the decision rule

$$f(\phi(x)) = \sum_{i=1}^{\ell} y_i \hat{\alpha}_i \hat{K}(\phi_i, \phi) + b. \tag{34}$$

Using SVM$_\Delta$+ algorithm described in Section 4, we can find the coefficients of expansion $\hat{\alpha}_i$ in (34). They are defined by the maximum (over $\hat{\alpha}$ and δ) of the functional

$$R(\hat{\alpha}, \delta) = \sum_{i=1}^{\ell} \hat{\alpha}_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \hat{\alpha}_i \hat{\alpha}_j \hat{K}(\phi_i, \phi_j) - \frac{1}{2\gamma} \sum_{i,j=1}^{\ell} y_i y_j (\hat{\alpha}_i - \delta_i)(\hat{\alpha}_j - \delta_j) K^*(x_i^*, x_j^*)$$

subject to the equality constraints

$$\sum_{i=1}^{\ell} \hat{\alpha}_i y_i = 0, \quad \sum_{i=1}^{\ell} \hat{\alpha}_i = \sum_{i=1}^{\ell} \delta_i$$

and the inequality constraints

$$0 \le \hat{\alpha}_i \le \delta_i + \Delta C, \quad 0 \le \delta_i \le C, \quad i = 1, \ldots, \ell$$

(see Section 4).

5.2 General Form of Knowledge Transfer

One can use many different ideas to represent knowledge obtained in space X^*. The main factors of these representations are concepts of fundamental elements of the knowledge. They could be, for example, just the support vectors (if the number of support vectors is not too big) or coordinates (features) $x^{t*},\ t = 1, \ldots, d$ of d-dimensional privileged space X^* (if the number of these features not too big). In the latter case, the small number of fundamental elements of knowledge would be composed of features x^{*k} in the privileged space that can be then approximated by regression functions $\phi_k(x)$. In general, using

privileged information it is possible to try transfer set of useful features for rule in X^* space into their image in X space.

The space where depiction rule is constructed can contain both features of space X and new features defined by the regression functions. The example of knowledge transfer described further in subsection 5.5 is based on this approach.

In general, the idea is to specify small amount important feature in privileged space and then try to transfer them (say, using non-linear regression technique) in decision space to construct useful (additional) features in decision space.

Note that in SVM framework, with the quadratic kernel the minimal number m of fundamental elements (features) does not exceed the dimensionality of space X^* (often, m is much smaller than dimensionality. This was demonstrated in multiple experiments with digit recognition by Burges 1996): in order to generate the same level of accuracy of the solution, it was sufficient to use m elements, where the value of m was at least 20 times smaller than the corresponding number of support vectors.

5.3 Kernels Involved in Intelligent Learning

In this paper, among many possible Mercer kernels (positive semi-definite functions), we consider the following three types:

1. Radial Basis Function (RBF) kernel:

$$K_{RBF_\sigma}(x, y) = \exp\{-\sigma^2(x - y)^2\}.$$

2. INK-spline kernel. Kernel for spline of order zero with infinite number of knots is defined as

$$K_{INK_0}(x, y) = \prod_{k=1}^{d}(\min(x^k, y^k) + \delta)$$

(δ is a free parameter) and kernel of spline of order one with infinite number of knots is defined in the non-negative domain and has the form

$$K_{INK_1}(x, y) = \prod_{k=1}^{d}\left(\delta + x^k y^k + \frac{|x^k - y^k| \min\{x_k, y^k\}}{2} + \frac{(\min\{x^k, y^k\})^3}{3}\right)$$

where $x^k \geq 0$ and $y^k \geq 0$ are k coordinates of d-dimensional vector x.

3. Homogeneous quadratic kernel

$$K_{Pol_2} = (x, y)^2,$$

where (x, y) is the inner product of vectors x and y.

The RBF kernel has a free parameter $\sigma > 0$; two other kernels have no free parameters. That was achieved by fixing a parameter in more general sets of functions: the degree of polynomial was chosen to be 2, and the order of INK-splines was chosen to be 1.

It is easy to introduce kernels for any degree of polynomials and any order of INK-splines. Experiments show excellent properties of these three types of kernels for solving many machine learning problems. These kernels also can be recommended for methods that use both mechanisms of Teacher-Student interaction.

5.4 Knowledge Transfer for Statistical Inference Problems

The idea of privileged information and knowledge transfer can be also extended to Statistical Inference problems considered in Vapnik and Izmailov (2015a) and Vapnik et al. (2015).

For simplicity, consider the problem of estimation[14] of conditional probability $P(y|x)$ from iid data

$$(x_1, y_1), ..., (x_\ell, y_\ell), \quad x \in X,\ y \in \{0, 1\}, \tag{35}$$

where vector $x \in X$ is generated by a fixed but unknown distribution function $P(x)$ and binary value $y \in \{0, 1\}$ is generated by an unknown conditional probability function $P(y = 1|x)$ (similarly, $P(y = 0|x) = 1 - P(y = 1|x)$); this is the function we would like to estimate.

As shown in Vapnik and Izmailov (2015a) and Vapnik et al. (2015), this requires solving the Fredholm integral equation

$$\int \theta(x - t)P(y = 1|t)dP(t) = P(y = 1, x),$$

where probability functions $P(y = 1, x)$ and $P(x)$ are unknown but iid data (35) generated according to joint distribution $P(y, x)$ are given. Vapnik and Izmailov (2015a) and Vapnik et al. (2015) describe methods for solving this problem, producing the solution

$$P_\ell(y = 1|x) = P(y = 1|x; (x_1, y_1), ..., (x_\ell, y_\ell)).$$

In this section, we generalize classical Statistical Inference problem of conditional probability estimation to a new model of Statistical Inference with Privileged Information. In this model, along with information defined in the space X, one has the information defined in the space X^*.

Consider privileged space X^* along with space X . Suppose that any vector $x_i \in X$ has its image $x_i^* \in X^*$. Consider iid triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell) \tag{36}$$

that are generated according to a fixed but unknown distribution function
$P(x, x^*, y)$. Suppose that, for any triplet (x_i, x_i^*, y_i), there exist conditional probabilities $P(y_i|x_i^*)$ and $P(y_i|x_i)$. Also, suppose that the conditional probability function $P(y|x^*)$, defined in the privileged space X^*, is *better* than the conditional probability function $P(y|x)$, defined in space X; here by "better" we mean that the *conditional entropy* for $P(y|x^*)$ is smaller than conditional entropy for $P(y|x)$:

$$-\int [\log_2 P(y = 1|x^*) + \log_2 P(y = 0|x^*)]\, dP(x^*) <$$

$$-\int [\log_2 P(y = 1|x) + \log_2 P(y = 0|x)]\, dP(x).$$

Our goal is to use triplets (36) for estimating the conditional probability
$P(y|x; (x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell))$ in space X better than it can be done with training pairs (35). That is, our goal is to find such a function

$$P_\ell(y = 1|x) = P(y = 1|x; (x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y))$$

14. The same method can be applied to all the problems described in Vapnik and Izmailov (2015a) and Vapnik et al. (2015).

that the following inequality holds:

$$-\int [\log_2 P(y=1|x;(x_i,x_i^*,y_i)_1^\ell) + \log_2 P(y=0|x;(x_i,x_i^*,y_i)_1^\ell)]dP(x) <$$

$$-\int [\log_2 P(y=1|x;(x_i,y_i)_1^\ell) + \log_2 P(y=0|x;(x_i,y_i)_1^\ell,)]dP(x).$$

Consider the following solution for this problem:

1. Using kernel $K(u^*, v^*)$, the training pairs (x_i^*, y_i) extracted from given training triplets (36) and the methods of solving our integral equation described in Vapnik and Izmailov (2015a) and Vapnik et al. (2015), find the solution of the problem in space of privileged information X^*:

$$P(y=1|x^*;(x_i^*,y_i)_1^\ell) = \sum_{i=1}^{\ell} \hat{\alpha}_i K(x_i^*, x^*) + b.$$

2. Find the fundamental elements of knowledge: vectors $u_1^*, ..., u_m^*$.

3. Using some universal kernels (say RBF or INK-Spline), find in the space X the approximations $\phi_k(x), k = 1, \ldots, m$ of the frames $(u_k^*, x^*)^2,\ k = 1, ..., m$.

4. Find the solution of the conditional probability estimation problem $P(y|\phi;(\phi_i,y_i)_1^\ell)$ in the space of pairs (ϕ, y), where $\phi = (\phi_1(x), \ldots, \phi_m(x))$.

5.5 Example of Knowledge Transfer Using Privileged Information

In this subsection, we describe an example where privileged information was used in the knowledge transfer framework. In this example, using set of of pre-processed video snapshots of a terrain, one has to separate pictures with specific targets on it (class $+1$) from pictures where there are no such targets (class -1).

The original videos were made using aerial cameras of different resolutions: a low resolution camera with wide view (capable to cover large areas quickly) and a high resolution camera with narrow view (covering smaller areas and thus unsuitable for fast coverage of terrain). The goal was to make judgments about presence or absence of targets using wide view camera that could quickly span large surface areas. The narrow view camera could be used during training phase for zooming in the areas where target presence was suspected, but it was not to be used during actual operation of the monitoring system, i.e., during test phase. Thus, the wide view camera with low resolution corresponds to standard information (space X), whereas the narrow view camera with high resolution corresponds to privileged information (space X^*).

The features for both standard and privileged information spaces were computed separately, using different specialized video processing algorithms, yielding 15 features for decision space X and 116 features for space of privileged information X^*.

The classification decision rules for presence or absence of targets were constructed using respectively,

- SVM with RBF kernel trained on 15 features of space X;

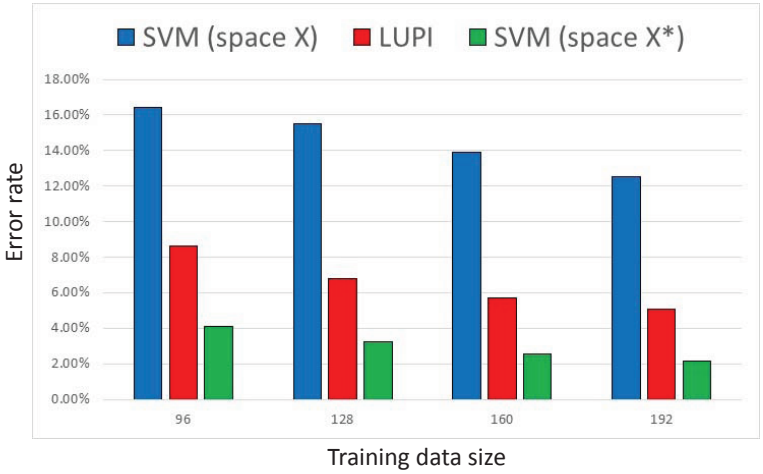

Figure 1: Comparison of SVM and knowledge transfer error rates: video snapshots example.

- SVM with RBF kernel trained on 116 features of space X^*;
- SVM with RBF kernel trained 15 original features of space X augmented with 116 knowledge transfer features, each constructed using regressions on the 15-dimensional decision space X (as outlined in subsection 5.2).

Parameters for SVMs with RBF kernel were selected using standard grid search with 6-fold cross validation.

Figure 1 illustrates performance (defined as an overage of error rate) of three algorithms each trained of 50 randomly selected subsets of sizes 64, 96, 128, 160, and 192: SVM in space X, SVM in space X^*, and SVM in space with transferred knowledge.

Figure 1 shows that, the larger is the training size, the better is the effect of knowledge transfer. For the largest training size considered in this experiment, the knowledge transfer was capable to recover almost 70% of the error rate gap between the error rates of SVM using only standard features and SVM using privileged features. In this Figure, one also can see that, even in the best case, the error rate using SVM in the space of privileged information is half of that of SVM in the space of transferred knowledge. This gap, probably, can be reduced even further by better selection of the fundamental concepts of knowledge in the space of privileged information and / or by constructing better regression.

5.6 General Remarks about Knowledge Transfer

5.6.1 What Knowledge Does Teacher Transfer?

In previous sections, we linked the knowledge of Intelligent Teacher about the problem of interest in X space to his knowledge about this problem in X^* space[15].

15. This two space learning paradigm with knowledge transfer for one space to another space reminds Plato's idea about *space of Ideas and space of Things* with transfer of knowledge from one space to another. This idea in different forms was explored by many philosophers.

One can give the following general mathematical justification for our model of knowledge transfer. Teacher knows that the goal of Student is to construct a good rule in space X with one of the functions from the set $f(x, \alpha),\ x \in X,\ \alpha \in \Lambda$ with capacity VC_X. Teacher also knows that there exists a rule of the same quality in space X^* – a rule that belongs to the set $f^*(x^*, \alpha^*),\ x^* \in X^*,\ \alpha^* \in \Lambda^*$ and that has a much smaller capacity VC_{X^*}. This knowledge can be defined by the ratio of the capacities

$$\kappa = \frac{VC_X}{VC_{X^*}}.$$

The larger is κ, the more knowledge Teacher can transfer to Student; also the larger is κ, the fewer examples will Student need to select a good classification rule.

5.6.2 Learning from Multiple Intelligent Teachers

Model of learning with Intelligent Teachers can be generalized for the situation when Student has $m > 1$ Intelligent Teachers that produce m training triplets

$$(x_{k_1}, x^{k*}_{k_1}, y_1), ..., (x_{k_\ell}, x^{k*}_{k_\ell}, y_\ell),$$

where $x_{k_t}, k = 1, ..., m,\ t = 1, ..., \ell$ are elements x of different training data generated by the same generator $P(x)$ and $x^{k*}_{k_t}, k = 1, ..., m,\ t = 1, ..., \ell$ are elements of the privileged information generated by kth Intelligent Teacher that uses generator $P_k(x^{k*}|x)$. In this situation, the method of knowledge transfer described above can be expanded in space X to include the knowledge delivered by all m Teachers.

5.6.3 Quadratic Kernel

In the method of knowledge transfer, the special role belongs to the quadratic kernel $(x_1, x_2)^2$. Formally, only two kernels are amenable for simple methods of finding the smallest number of fundamental elements of knowledge: the linear kernel (x_1, x_2) and the quadratic kernel $(x_1, x_2)^2$.

Indeed, if linear kernel is used, one constructs the separating hyperplane in the space of privileged information X^*

$$y = (w^*, x^*) + b^*,$$

where vector of coefficients w^* also belongs to the space X^*, so there is only one fundamental element of knowledge, i.e., the vector w^*. In this situation, the problem of constructing the regression function $y = \phi(x)$ from data

$$(x_1, (w^*, x^*_1)), ..., (x_\ell, (w^*, x^*_\ell)) \tag{37}$$

has, generally speaking, the same level of complexity as the standard problem of pattern recognition in space X using data (35). Therefore, one should not expect performance improvement when transferring the knowledge using (37).

With quadratic kernel, one obtains fewer than d fundamental elements of knowledge in d-dimensional space X^* (experiments show that the number of fundamental elements can be significantly smaller than d). According to the methods described above, one defines the knowledge in space X^* as a linear combination of m frames. That is, one splits the desired

function into m fragments (a linear combination of which defines the decision rule) and then estimates each of m functions $\phi_k(x)$ separately, using training sets of size ℓ. The idea is that, in order to estimate a fragment of the knowledge well, one can use a set of functions with a smaller capacity than is needed to estimate the entire function $y = f(x),\ x \in X$. Here privileged information can improve accuracy of estimation of the desired function.

To our knowledge, there exists only one nonlinear kernel (the quadratic kernel) that leads to an exact solution of the problem of finding the fundamental elements of knowledge. For all other nonlinear kernels, the problems of finding the minimal number of fundamental elements require difficult (heuristic) computational procedures.

6. Conclusions

In this paper, we tried to understand mechanisms of learning that go beyond brute force methods of function estimation. In order to accomplish this, we used the concept of Intelligent Teacher who generates privileged information during training session. We also described two mechanisms that can be used to accelerate the learning process:

1. The mechanism to control Student's concept of similarity between training examples.

2. The mechanism to transfer knowledge from the space of privileged information to the desired decision rule.

It is quite possible that there exist more mechanisms in Teacher-Student interactions and thus it is important to find them.

The idea of privileged information can be generalized to any statistical inference problem creating non-symmetric (two spaces) approach in statistics.

Teacher-Student interaction constitutes one of the key factors of intelligent behavior and it can be viewed as a basic element in understanding intelligence (for both machines and humans).

Acknowledgments

This material is based upon work partially supported by AFRL and DARPA under contract FA8750-14-C-0008. Any opinions, findings and / or conclusions in this material are those of the authors and do not necessarily reflect the views of AFRL and DARPA.

We thank Professor Cherkassky, Professor Gammerman, and Professor Vovk for their helpful comments on this paper.

References

R. Brachman and H. Levesque. *Knowledge Representation and Reasoning*. Morgan Kaufman Publishers, San Francisco, CA, 2004.

C. Burges. Simplified support vector decision rules. In *13th International Conference on Machine Learning, Proceedings*, pages 71–77, 1996.

A. Chervonenkis. *Computer Data Analysis (in Russian)*. Yandex, Moscow, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of mathematics : stochastic modelling and applied probability. Springer, 1996.

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in banach spaces. *Theoretical Computer Science*, 261(1):81–90, 2001.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

L. Liang and V. Cherkassky. Connection between SVM+ and multi-task learning. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, 2008*, pages 2048–2054, 2008.

D. Pechyony, R. Izmailov, A. Vashist, and V. Vapnik. Smo-style algorithms for learning using privileged information. In *International Conference on Data Mining*, pages 235–241, 2010.

B. Ribeiro, C. Silva, N. Chen, A. Vieira, and J. das Neves. Enhanced default risk models with svm+. *Expert Systems with Applications*, 39(11):10140–10152, 2012.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pages 416–426, London, UK, UK, 2001. Springer-Verlag.

V. Sharmanska, N. Quadrianto, and C. Lampert. Learning to rank using privileged information. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 825–832. IEEE, 2013.

V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., 1982.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer–Verlag, 2nd edition, 2006.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition (in Russian)*. Nauka, Moscow, 1974.

V. Vapnik and R. Izmailov. Statistical inference problems and their rigorous solutions. In Alexander Gammerman, Vladimir Vovk, and Harris Papadopoulos, editors, *Statistical Learning and Data Sciences*, volume 9047 of *Lecture Notes in Computer Science*, pages 33–71. Springer International Publishing, 2015a.

V. Vapnik and R. Izmailov. Learning with intelligent teacher: Similarity control and knowledge transfer. In A. Gammerman, V. Vovk, and H. Papadopoulos, editors, *Statistical Learning and Data Sciences*, volume 9047 of *Lecture Notes in Computer Science*, pages 3–32. Springer International Publishing, 2015b.

V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

V. Vapnik, I. Braga, and R. Izmailov. Constructive setting for problems of density ratio estimation. *Statistical Analysis and Data Mining*, 8(3):137–146, 2015.