Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer

In memory of Alexey Chervonenkis

Vladimir Vapnik[1,2(✉)] and Rauf Izmailov[3]

[1] Columbia University, New York, NY, USA
vladimir.vapnik@gmail.com
[2] AI Research Lab, Facebook, New York, NY, USA
[3] Applied Communication Sciences, Basking Ridge, NJ, USA
rizmailov@appcomsci.com

Abstract. This paper introduces an advanced setting of machine learning problem in which an Intelligent Teacher is involved. During training stage, Intelligent Teacher provides Student with information that contains, along with classification of each example, additional privileged information (explanation) of this example. The paper describes two mechanisms that can be used for significantly accelerating the speed of Student's training: (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 [21], [18], [19], [5]) 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 of Structural Risk Minimization (SRM), which always achieves the smallest risk.
- The effective algorithms (such as Support Vector Machines (SVM)), that realize the consistency property of SRM principle.

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.

A. Gammerman et al. (Eds.): SLDS 2015, LNAI 9047, pp. 3–32, 2015.
DOI: 10.1007/978-3-319-17091-6_1

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[1] 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 are also based on special mechanisms that can significantly accelerate the learning process. In order for a learning machine to use fewer observations, it has to use these mechanisms as well.

This paper considers the model of learning that includes 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 this 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.*
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.*

[1] Japanese proverb assesses teacher's influence as follows: "better than a thousand days of diligent study is one day with a great teacher."

The first mechanism [20] was introduced in 2006, and here we are mostly reproduce results obtained in [22]. The second mechanism is introduced in this paper for the first time.

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 Model 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 of training example generated by Nature, Intelligent Teacher generates both its label y_i and the privileged information x_i^* using some unknown conditional probability function $P(x_i^*, y_i|x_i)$.

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 past patients.

However, for past 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 images x along with the pathologist's reports x^* as a privileged information in order to make a better classification rule just in pixel space X: classification by a pathologist is a difficult and time-consuming procedure, so fast decisions during surgery should be made automatically, without consulting a pathologist.

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 the 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 all machine learning problems.

In Section 4, we describe a mechanism that allows one to take advantage of privileged information by controlling Student's concepts of similarity between training examples. However, we first describe statistical properties enabling the use of privileged information.

3 Statistical Analysis of the Rate of Convergence

According to the bounds developed in the VC theory [21], [19], 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 $f(x, \alpha), \alpha \in \Lambda$ that separates the training data (4) without errors:

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

In this case, the function $f(x, \alpha_\ell)$ that minimizes the empirical risk (on training set (4)) with probability $1 - \eta$ has the bound

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

where $p(yf(x, \alpha_\ell) \leq 0)$ is the probability of error for the function $f(x, \alpha_\ell)$ and h is 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 $f(x, \alpha), \ \alpha \in \Lambda$ 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 the 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[2] 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.

In the separable case, SVM constructs (in space Z which we, for simplicity, 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

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

[2] The VC theory also gives more accurate estimate of the rate of convergence; however, the difference remains essentially the same.

while 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

$$(y_i(w, z_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, ..., \ell.$$

That is, in the separable case, SVM uses ℓ observations for estimation of N coordinates of vector w, while 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, Teacher supplies triplets

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

where ξ_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 maximize 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 [22].

Proposition 1. *Let $f(x, \alpha_0)$ be a function from the set of indicator functions $f(x, \alpha), \alpha \in \Lambda$ 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:

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

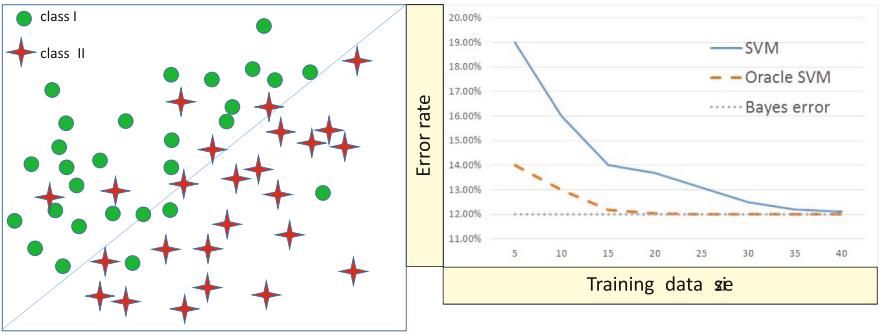

Fig. 1. Comparison of Oracle SVM and standard SVM

That is, for Oracle SVM, the rate of convergence is $1/\ell$ even in the non-separable case. Figure 1 illustrates this: the left half of the figure shows synthetic data for a binary classification problem using the set of linear rules with Bayesian rule having error rate 12% (the diagonal), while the right half of the figure illustrates the rates of convergence for standard SVM and Oracle SVM. While both converge to the Bayesian solution, Oracle SVM does it much faster.

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[3] are selected;
3. generate privileged information for training examples supplying Student, instead of pairs (4), with triplets

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

[3] 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 dimension, respectively. Here we consider two sets with finite VC dimensions.

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 $f^*(x^*, \alpha^*)$. 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 $f(x, \alpha_\ell)$ be a function that solves this optimization problem. For this function, the following proposition holds true [22].

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 $f^*(x^*, \alpha^*), \alpha^* \in \Lambda^*$ 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 $f^*(x^*, \alpha^*), \alpha^* \in \Lambda^*$ (for example, those that satisfy the conditions defined by Tsybakov [15] or the conditions defined by Steinwart and Scovel [17]), the convergence can be faster (as $O([1/\ell]^\delta)$, $\delta > 1/2$).

A well-selected privileged information space X^* and Teacher's explanation $P(x^*, y|x)$ 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 $P(x^*, y|x)$, 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 SVM+ for Similarity Control in LUPI Paradigm

In this section, we extend SVM to the method called SVM+, which allows one to solve machine learning problems in the LUPI paradigm [22].

Consider again 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.[4]

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

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

where w^* is an element in Z^*. In SVM+, the goal is to minimize the functional

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

subject to the linear constraints

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

where $[u]_+ = \max\{0, u\}$.

The structure of this problem mirrors the structure of the primal problem for standard SVM, while containing one additional parameter $\gamma > 0$.

To find the solution of this optimization problem, we use the equivalent setting: we minimize the functional

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

subject to constraints

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

and

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

[4] In [22], the case of privileged information being available only for a subset of examples is considered: specifically, for examples with non-zero values of slack variables.

and

$$\zeta_i \geq 0, \quad \forall i = 1, ..., \ell. \tag{11}$$

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

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

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

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

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 $\alpha_i, \beta_i, \nu_i,\ i = 1, ..., \ell$).

The necessary conditions for minimum of (12) are

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

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

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

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

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

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

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

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

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

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

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

$$\alpha_i \ge 0, \quad i = 1, ..., \ell \tag{21}$$

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

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

and to the slack function

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

The Karush-Kuhn-Tacker conditions for this problem are

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

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 $\alpha_k^0 \ne 0$ and $\delta_k^0 \ne C$.

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+ 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

$$\mathcal{L}(\alpha, \delta) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \frac{1}{2\gamma} \sum_{i,j=1}^{\ell} (\alpha_i - \delta_i)(\alpha_j - \delta_j) K^*(x_i^*, x_j^*) \tag{22}$$

subject to constraints (18) – (21).

Remark. In the special case $\delta_i = \alpha_i$, our optimization problem becomes equivalent to 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+ and SVM solutions is defined by the last term in objective function (22). 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+ solution is defined by objective function (22) that uses two expressions of similarities between observations: one (x_i and x_j) that comes from space X and another one (x_i^* and x_j^*) that comes from space of privileged information X^*. That is, Intelligent Teacher changes the optimal solution by correcting concepts of similarity.

The last term in equation (22) defines the instrument for Intelligent Teacher to control the concept of similarity of Student.

To find value of b, one has to find a sample (x_k, x_k^*, y_k) for which $\alpha_k > 0,\ \delta_k < C$ and compute

$$b = 1 - y_k \left[\sum_{i=1}^{\ell} y_i \alpha_i K(x_i, x_k) \right].$$

Efficient computational implementation of this SVM+ algorithm for classification and its extension for regression can be found in [14] and [22], respectively.

5 Three Examples of Similarity Control Using Privileged Information

In this section, we describe three different types of privileged information (advanced technical model, future events, holistic description), used in similarity control setting [22].

5.1 Advanced Technical Model as Privileged Information

Homology classification of proteins is a hard problem in bioinformatics. Experts usually rely on hierarchical schemes leveraging molecular 3D-structures, which are expensive and time-consuming (if at all possible) to obtain. The alternative information on amino-acid sequences of proteins can be collected relatively easily, but its correlation with 3D-level homology is often poor (see Figure 2). The practical problem is thus to construct a rule for classification of proteins based on their amino-acid sequences as standard information, while using available molecular 3D-structures as privileged information.

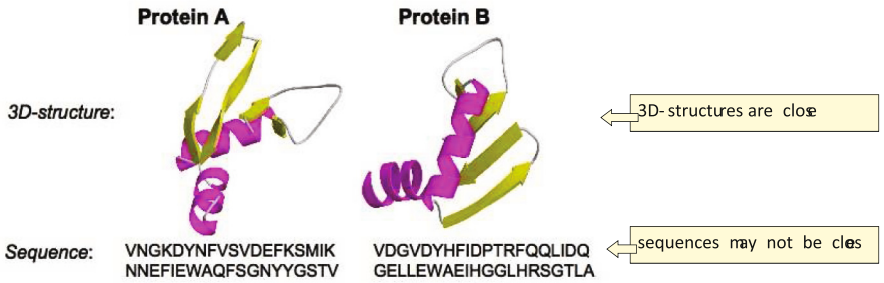

Fig. 2. 3D-structures and amino-acid sequences of proteins

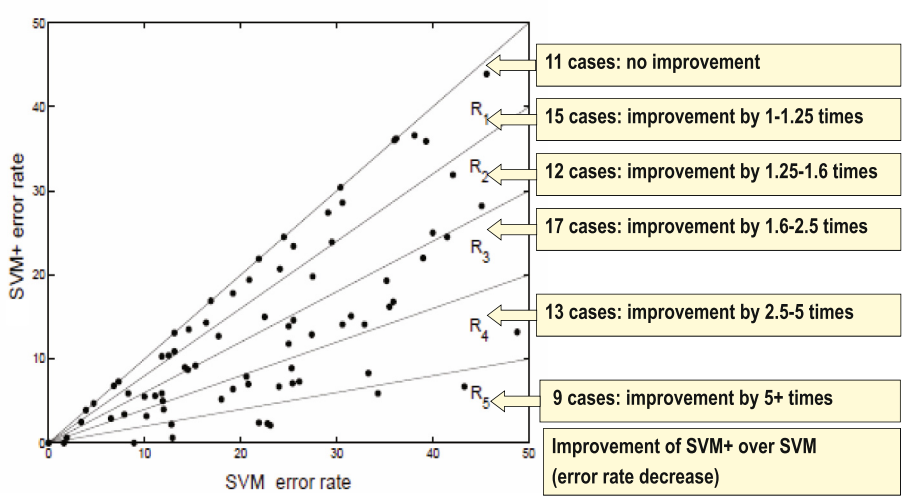

Fig. 3. Comparison of SVM and SVM+ error rates

Since SVM has been successfully used [8], [9] to construct protein classification rules based on amino-acid sequences, the natural next step was to see what performance improvement can be obtained by using 3D-structures as privileged information and applying SVM+ method of similarity control. The experiments used SCOP (Structural Classification of Proteins) database [11], containing amino-acid sequences and their hierarchical organization, and PDB (Protein Data Bank) [2], containing 3D-structures for SCOP sequences. The classification goal was to determine homology based on protein amino-acid sequences from 80 superfamilies (3rd level of hierarchy) with the largest number of sequences. Similarity between amino-acid sequences (standard space) and 3D-structures (privileged space) was computed using the *profile-kernel* [8] and *MAMMOTH* [13], respectively.

Standard SVM classification based on 3D molecular structure had an error rate smaller than 5% for almost any of 80 problems, while SVM classification using protein sequences gave much worse results (in some cases, the error rate was up to 40%).

Figure 3 displays comparison of SVM and SVM+ with 3D privileged information. It shows that SVM+ never performed worse than SVM. In 11 cases it gave exactly the same result, while in 22 cases its error was reduced by more than 2.5 times. Why does the performance vary so much? The answer lies in the nature of the problem. For example, both diamond and graphite consist of the same chemical element, carbon, but they have different molecular structures. Therefore, one can only tell them apart using their 3D structures.

5.2 Future Events as Privileged Information

Time series prediction is used in many statistical applications: given historical information about the values of time series up to moment t, predict the value (qualitative setting) or the deviation direction (positive or negative; quantitative setting) at the moment $t + \Delta$.

One of benchmark time series for prediction algorithms is the quasi-chaotic (and thus difficult to predict) Mackey-Glass time series, which is the solution of the equation [10], [4]

$$\frac{dx(t)}{dt} = -ax(t) + \frac{bx(t-\tau)}{1 + x^{10}(t-\tau)}.$$

Here a, b, and τ (delay) are parameters, usually assigned the values $a = 0.1,\ b = 0.2,\ \tau = 17$ with initial condition $x(\tau) = 0.9$.

The qualitative prediction setting (will the future value $x(t+T)$ be larger or smaller than the current value $x(t)$?) for several lookahead values of T (specifically, $T = 1,\ T = 5,\ T = 8$) was used for comparing the error rates of SVM and SVM+. The standard information was the vector $x_t = (x(t-3), x(t-2), x(t-1), x(t))$ of current observation and three previous ones, whereas the privileged information was the vector $x_t^* = (x(t+T-2), x(t+T-1), x(t+T+1), x(t+T+2))$ of four future events.

The experiments covered various training sizes (from 100 to 500) and several values of T (namely, $T = 1$, $T = 5$, and $T = 8$). In all the experiments, SVM+ consistently outperformed SVM, with margin of improvement being anywhere between 30% and 60%; here the margin was defined as relative improvement of error rate as compared to the (unattainable) performance of the specially constructed Oracle SVM.

5.3 Holistic Description as Privileged Information

This example is an important one, since holistic privileged information is most frequently used by Intelligent Teacher. In this example, we consider the problem of classifying images of digits 5 and 8 in the MNIST database. This database

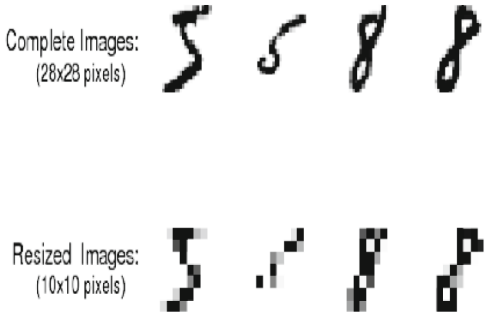

Fig. 4. Sample MNIST digits ans their resized images

contains digits as 28*28 pixel images; there are 5,522 and 5,652 images of digits 5 and 8, respectively. Distinguishing between these two digits in 28*28 pixel space is an easy problem. To make it more challenging, the images were resized to 10*10 pixels (examples are shown in Figure 4). A hundred examples of 10*10 images were randomly selected as a training set, another 4,000 images were used as a validation set (for tuning the parameters in SVM and SVM+) and the remaining 1,866 images constituted the test set.

For every training image, its holistic description was created (using natural language). For example, the first image of 5 (see Figure 4) was described as follows:

> *Not absolute two-part creature. Looks more like one impulse. As for two-partness the head is a sharp tool and the bottom is round and flexible. As for tools it is a man with a spear ready to throw it. Or a man is shooting an arrow. He is firing the bazooka. He swung his arm, he drew back his arm and is ready to strike. He is running. He is flying. He is looking ahead. He is swift. He is throwing a spear ahead. He is dangerous. It is slanted to the right. Good snaked-ness. The snake is attacking. It is going to jump and bite. It is free and absolutely open to anything. It shows itself, no kidding. Its bottom only slightly (one point!) is on earth. He is a sportsman and in the process of training. The straight arrow and the smooth flexible body. This creature is contradictory - angular part and slightly roundish part. The lashing whip (the rope with a handle). A toe with a handle. It is an outside creature, not inside. Everything is finite and open. Two open pockets, two available holes, two containers. A piece of rope with a handle. Rather thick. No loops, no saltire. No hill at all. Asymmetrical. No curlings.*

The first image of 8 (Figure 4) was described as follows:

> *Two-part creature. Not very perfect infinite way. It has a deadlock, a blind alley. There is a small right-hand head appendix, a small shoot. The right-hand appendix. Two parts. A bit disproportionate. Almost equal. The upper*

one should be a bit smaller. The starboard list is quite right. It is normal like it should be. The lower part is not very steady. This creature has a big head and too small bottom for this head. It is nice in general but not very self-assured. A rope with two loops which do not meet well. There is a small upper right-hand tail. It does not look very neat. The rope is rather good - not very old, not very thin, not very thick. It is rather like it should be. The sleeping snake which did not hide the end of its tail. The rings are not very round - oblong - rather thin oblong. It is calm. Standing. Criss-cross. The criss-cross upper angle is rather sharp. Two criss-cross angles are equal. If a tool it is a lasso. Closed absolutely. Not quite symmetrical (due to the horn).

These holistic descriptions were mapped into 21-dimensional feature vectors. Examples of these features (with range of possible values) are: `two-part-ness` (0 - 5); `tilting to the right` (0 - 3); `aggressiveness` (0 - 2); `stability` (0 - 3); `uniformity` (0 - 3), and so on. The values of these features (in the order they appear above) for the first 5 and 8 are [2, 1, 2, 0, 1], and [4, 1, 1, 0, 2], respectively. Holistic descriptions and their mappings were created prior to the learning process by an independent expert; all the datasets are publicly available at [12].

The goal was to construct a decision rule for classifying 10*10 pixel images in the 100-dimensional standard pixel space X and to leverage the corresponding 21-dimensional vectors as the privileged space X^*. This idea was realized using the SVM+ algorithm described in Section 4. For every training data size, 12 different random samples selected from the training data were used and the average of test errors was calculated.

To understand how much information is contained in holistic descriptions, 28*28 pixel digits (784-dimensional space) were used instead of the 21-dimensional holistic descriptions in SVM+ (the results shown in Figure 5). In this setting, when using 28*28 pixel description of digits, SVM+ performs worse than SVM+ using holistic descriptions.

6 Transfer of Knowledge Obtained in Privileged Information Space to Decision Space

In this section, we consider one of the most important mechanisms of Teacher-Student interaction: using privileged information to transfer knowledge from Teacher to Student.

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*.

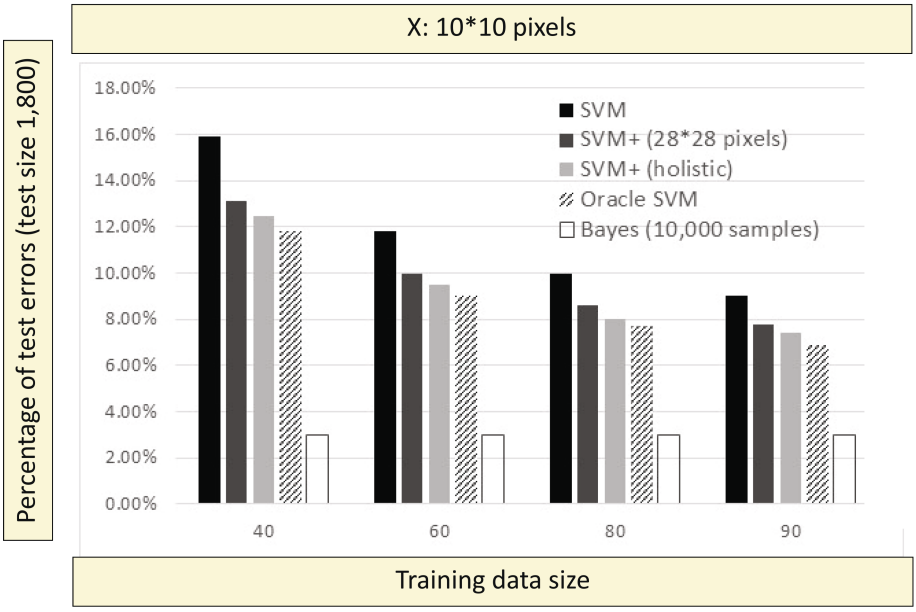

Fig. 5. Error rates for the digit recognition task

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) which 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 s process; Teacher can holistically 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 is developing a special language that is appropriate for description x_i^* of cancer development using the model he believes in. Using this language, Teacher supplies Student with privileged information x_i^* for the image x_i by generating training triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell). \tag{23}$$

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: the set of admissible functions in space X has to be rich (has a large VC dimension), while the 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 (23):

1. The problem of constructing a rule $y = f(x)$ for classification of biopsy in the pixel space X using data

$$(x_1, y_1), ..., (x_\ell, y_\ell). \tag{24}$$

2. The problem of constructing a rule $y = f^*(x^*)$ for classification of biopsy in the space X^* using data

$$(x_1^*, y_1), ..., (x_\ell^*, y_\ell). \tag{25}$$

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.[5]

Since 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 error rate for rule $y = f_\ell^*(x^*)$ in X^* will be better[6] than those for the rule $y = f_\ell(x)$ in X. That is, generally speaking, the classification rule $y = f_\ell^*(x^*)$ will be more accurate than classification rule $y = f_\ell(x)$.

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?

6.1 Knowledge Representation

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 [7], [16], such rule has the form

$$f_\ell^*(x^*) = \sum_{i=1}^{\ell} \gamma_i K^*(x_i^*, x^*) + b, \tag{26}$$

[5] The rule constructed in space X^* cannot be better than the best possible rule in space X, since all information originates in space X.

[6] According to VC theory, the guaranteed bound on accuracy of the chosen rule depends only on two factors: frequency of errors on training set and VC dimension of admissible set of functions.

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

Suppose that, using data (25), we found a good rule (26) with coefficients $\gamma_i = \gamma_i^*,\ i = 1, ..., \ell$ and $b = b^*$. This is now our knowledge about our classification problem. Let us formalize the description of this knowledge.

Consider three elements of knowledge representation used in AI [1]:

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* the smallest number of the vectors $u_1^*, ..., u_m^*$ from space X^* that can approximate[7] the main part of the rule (26):

$$f_\ell^*(x^*) - b = \sum_{i=1}^{\ell} \gamma_i^* K^*(x_i^*, x^*) \approx \sum_{k=1}^{m} \beta_k^* K^*(u_k^*, x^*). \tag{27}$$

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.

6.2 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

$$\phi_k(x_i) \approx K^*(u_k^*, x_i^*) \tag{28}$$

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

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 (27) 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, our hope is that transfer of m fragments of knowledge from space X^* into space X can be done with higher accuracy than estimating function $y = f_\ell(x)$ from data (24).

[7] In machine learning, they are called the reduced number of support vectors [3].

After finding approximation 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 $\delta_k = \beta_k^*$ (taken from (26)) if approximations (28) are accurate. Otherwise, coefficients δ_k can be estimated from the training data, as shown in Section 6.3.

Finding 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 fundamental elements of knowledge, we have to minimize (over vectors $u_1^*, ..., u_m^*$ and values $\beta_1, ..., \beta_m$) the functional

$$R(u_1^*, ..., u_m^*; \beta_1, ..., \beta_m) = \tag{29}$$

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

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

The last equality was derived from the following property of the inner product for functions from RKHS [7], [16]:

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

Fundamental Elements of Knowledge for Homogenous Quadratic Kernel. For general kernel functions $K^*(\cdot, \cdot)$, minimization of (29) is a difficult computational problem. However, for the special homogenous quadratic kernel

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

this problem has a simple exact solution [3]. For this kernel, we have

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

Let us look for solution in set of orthonormal vectors $u_i^*, ..., u_m^*$ for which we can rewrite (30) as follows

$$\hat{R} = \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i^*, x_j^*)^2 - 2 \sum_{i=1}^{\ell} \sum_{j=1}^{m} y_i \alpha_i \beta_s (x_i^*, u_s^*)^2 + \sum_{s=1}^{m} \beta_s^2 (u_s^*, u_s^*)^2. \tag{31}$$

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

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

Introducing notation

$$S = \sum_{i=1}^{\ell} y_i \alpha_i x_i^* x^{*T}, \tag{32}$$

we conclude that the solutions satisfy the equation

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

That is, the solutions $u_1^*, ..., u_m^*$ are the set of eigenvectors of the matrix S corresponding to non-zero eigenvalues $\beta_1, \ldots, \beta_m$, which are coefficients of expansion of the classification rule on the frames $(u_k, x^*)^2$, $k = 1, \ldots, m$.

Using (32), one can rewrite the functional (31) in the form

$$\hat{R} = \mathbf{1}^{\mathbf{T}} S \mathbf{1} - \sum_{k=1}^{m} \beta_k^2, \tag{33}$$

where we have denoted by $\mathbf{1}$ 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 m of eigenvectors corresponding to m eigenvalues with largest absolute values. One chooses such value of m that makes functional (33) small. The number m does not exceed n (the dimensionality of matrix S).

Finding Images of Frames in Space X. Let us call the conditional average 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

$$(x_1, K^*(u_k^*, x_1^*)), ..., (x_\ell, K^*(u_k^*, x_\ell^*)), \quad k = 1, \ldots, m, \tag{34}$$

where pairs (x_i, x_i^*) belong to elements of training triplets (23).

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 (34), one constructs the transformation of the space X into m-dimensional feature space

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

where k coordinates of vector function $\phi(x)$ are defined as $\phi_k = \phi_k(x)$.

6.3 Algorithms for Knowledge Transfer

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

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

hold true for almost all pairs (x_i, x_i^*) generated according to $P(x, x^*, y)$. 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

$$f(\phi(x)) = \sum_{i=1}^{\ell} y_i \hat{\alpha}_i \hat{K}(\phi_i, \phi) + b. \tag{35}$$

Using SVM+ algorithm described in Section 4, we can find the coefficients of expansion $\hat{\alpha}_i$ in (35). 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 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 inequality constraints

$$\hat{\alpha}_i \geq 0, \quad 0 \leq \delta_i \leq C, \quad i = 1, \ldots, \ell$$

(see Section 4).

Remark. One can use different ideas to represent knowledge obtained in the 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 or coordinates x^{t*}, $t = 1, ..., d$ of d-dimensional privileged space X^*. However, the fundamental elements defined above have some good properties: for the quadratic kernel, the number m of fundamental elements does not exceed the dimensionality of the space. Also, as was shown in multiple experiments with digit recognition [3], in order to generate the same level of accuracy of the solution, it was sufficient to use m elements, where the value m was at least 20 times smaller than the corresponding number of support vectors.

6.4 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 one with infinite number of knots. It is defined in the nonnegative domain and has the form

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

   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. Note that INK-splines are sensitive to the selection of minimum value a of coordinates x; as illustrated in [6], for reliable performance one should select $a = -3$ and reset all the values smaller than that to a (assuming all the coordinates are normalized to $N(0, 1)$ by subtracting the empirical means and dividing the values by empirical standard deviations).

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.

6.5 Knowledge Transfer for Statistical Inference Problems

The idea of privileged information and knowledge transfer can be also extended to Statistical Inference problems considered in [23], [24].

For simplicity, consider the problem of conditional probability $P(y|x)$ estimation[8] from iid data

$$(x_1, y_1), ..., (x_\ell, y_\ell), \quad x \in X,\ y \in \{0, 1\}, \tag{36}$$

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 $P(y = 1|x)$, $(P(y = 0|x) = 1 - P(y = 1|x))$; this is the function we would like to estimate.

As shown in [23], [24], this requires solving the Fredholm integral equation

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

if probability functions $P(y = 1, x)$ and $P(x)$ are unknown but iid data (36) generated according to joint distribution $P(y, x)$ are given. Papers [23], [24] 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

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell) \tag{37}$$

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"

[8] The same method can be applied to all problems described in [23], [24].

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 (37) 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 (36). That is, our goal is to find such a function

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

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 $K(\hat{x}^*, x^*) = (\hat{x}^*, x^*)^2$, the training pairs pairs (x_i^*, y_i) extracted from given training triplets (37) and the methods of solving our integral equation described in [23], [24], find the solution of the problem in space of privileged information X^*:

$$P(y=1|x^*; (x_i^*, y_i)_1^\ell) = \sum_{i=1}^{\ell} \hat{\alpha}_i (x_i^*, x^*)^2 + b.$$

2. Using matrix $S = \sum_{i=1}^{\ell} \hat{\alpha}_i x_i^* x_i^{*T}$, find the fundamental elements of knowledge (the eigenvectors $u_1^*, ..., u_m^*$ corresponding to the largest norm of eigenvalues $\beta_1, ..., \beta_m$ of matrix S).
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 $\phi = (\phi_1(x), \ldots, \phi_m(x))$.

6.6 General Remarks About Knowledge Transfer

What Knowledge Does Teacher Transfer? In previous sections, we linked the knowledge of Intelligent Teacher about the problem of interest in space X to his knowledge about this problem in space X^*. For knowledge transfer, one can

consider a more general model. Teacher knows that 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 in space X^* of the same quality which belongs to the set $f^*(x^*, \alpha^*),\ x^* \in X^*,\ \alpha^* \in \Lambda^*$ that has 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 and fewer examples will Student need to select a good classification rule.

What Are the Roots of Intelligence? In this paper, we defined the roots of intelligence via privileged information produced by Teacher according to Intelligent generator $P(x^*, y|x)$. Existence of triplets[9] (x, x^*, y) in description of the World events reflects *Holism* branch of philosophy, in which any event has multiple descriptions that cannot be reduced to a single point of view. In the opposite *Reductionism* branch of philosophy, descriptions of events can be reduced to a single major point of view (this is reflected in many branches of natural science).

We believe that Intelligent learning reflects Holism philosophy in particular, in multi-level interactions. It appears that intelligent learning is a multi-level representation of events and transfer elements of knowledge between different levels. We also believe that attempts to *significantly* improve the rate of learning in the classical setting using *only* more elaborate mathematical techniques are somewhat akin to Baron Munchausen's feat of pulling himself from a swamp[10]: for a significant improvement in learning process, Student needs *additional* (privileged) information in the same way the real-life Baron would have needed at least a solid foothold for getting out of his predicament.

Holistic Description and Culture. Generally speaking, Student and Teacher can have different sets of functions in Space X^*. In Section 5.3, we presented holistic descriptions of digits 5 and 8 as privileged information for training data reflecting impression of Prof. of Russian Poetry Natalia Pavlovitch. To transform her linguistic representation of privileged information into formal code, Prof. Pavlovitch suggested features for the transformation. One can see that transformation of privileged information from human language into code is very individual. Two different Students can obtain different functions as a result of transfer of the same knowledge given by Teacher. Therefore, in real life, Teacher has to elevate Student on the level of culture where Student can appropriately understand the privileged information.

[9] One can generalize the knowledge transfer method for multiple levels of privileged information, say, for two levels using quadruples (x, x^*, x^{**}, y).

[10] Baron Munchausen, the fictional character known for his tall tales, once pulled himself (and his horse) out of a swamp by his hair.

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 fundamental elements of knowledge (and therefore for knowledge representation): 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 – the vector w^*. In this situation, the problem of constructing the regression function $y = \phi(x)$ from data

$$(x_1, (w^*, x_1^*)), ..., (x_\ell, (w^*, x_\ell^*)) \tag{38}$$

has, generally speaking, the same level of complexity as the standard problem of pattern recognition in space X using data (36). Therefore, one should not expect performance improvement when transferring the knowledge using (38).

With quadratic kernel, one obtains $1 \leq m \leq d$ fundamental elements of knowledge in d-dimensional space X^*. In this situation, 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 (simplified) 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 well a fragment of the knowledge, 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 fundamental elements require difficult (heuristic) computational procedures.

Some Philosophical Interpretations. Classical German philosophy had formulated the following general idea (Hegel,1820):

> *What is reasonable*[11] *is real and what is real is reasonable.*

If we interpret the word "reasonable" as "has a good mathematical justification" and word "real" as "is realized by Nature" we can reformulate this idea as follows:

> Models that have a good mathematical justification are realized by the Nature and models that are realized by the Nature have a good mathematical justification.

[11] Some translations from German use "rational" instead of "reasonable".

From this point of view, the existence of only one kernel (the quadratic polynomial $(x_i, x_j)^2$) that allows one to find the exact solution of the knowledge representation problem means that there exists only one good level of structural complexity for privileged information that can be used by Intelligent Teacher: not too simple (such as based on linear kernel (x_1, x_2)), but not too complex (such as based on kernels $(x_1, x_2)^s,\ s > 2$). This claim, however, is not based on a proof, it is rather based on belief in classical German philosophy.

7 Conclusions

In this paper, we tried to understand mechanisms of human learning that go beyond brute force methods of function estimation. In order to accomplish this, we introduced the concept of Intelligent Teacher who generates privileged information during training session. We 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 [12] .

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 (in both machines and humans).

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

[12] In [23], we discuss the idea of replacing the SVM method with the so-called V-matrix method of conditional probability estimation. As was shown in Section 6.5, privileged information also can be used for estimation of the conditional probability function. However, for estimation of this function, Intelligent Teacher can also include in the privileged information his evaluation of the probability $p(x_i)$ that event x_i belongs to class $y = 1$ given his guess

$$a_i \le p(x_i) \le b_i,$$

where $0 \le a_i < b_i \le 1$ are some values. For example, in the problem of classification of a biopsy image x_i, the pathologist can give his assessment of cancer probability (hopefully non-trivial $(a_i > 0, b_i < 1)$). Footnote 10 in [23] shows how such information can be used for evaluation of conditional probability function.

References

1. Brachman, R., Levesque, H.: Knowledge Representation and Reasoning. Morgan Kaufmann, San Francisco (2004)
2. Berman, H., Westbrook, J., Feng, Z., Gillil, G., Bhat, T., Weissig, H., et al.: The protein data bank. Nucleic Acids Research **28**, 235–242 (2000)
3. Burges, C.: Simplified support vector decision rules. In: 13th International Conference on Machine Learning, pp. 71–77 (1996)
4. Casdagli, M.: Nonlinear prediction of chaotic time series. Physica D **35**, 335–356 (1989)
5. Chervonenkis, A.: Computer Data Analysis (in Russian). Yandex, Moscow (2013)
6. Izmailov, R., Vapnik, V., Vashist, A.: Multidimensional Splines with Infinite Number of Knots as SVM Kernels. In: International Joint Conference on Neural Networks, pp. 1096–1102. IEEE Press, New York (2013)
7. Kimeldorf, G., Wahba, G.: Some Results on Tchebycheffian Spline Functions. Journal of Mathematical Analysis and Applications **33**, 82–95 (1971)
8. Kuang, R., Le, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., et al.: Profile-Based String Kernels for Remote Homology Detection and Motif Extraction. Journal of Bioinformatics and Computational Biology **3**, 527–550 (2005)
9. Liao, L., Noble, W.: Combining Pairwise Sequence Similarity and Support Vector Machines for Remote Protein Homology Detection. Journal of Computational Biology **10**, 857–868 (2003)
10. Mukherjee, S., Osuna, E., Girosi, F.: Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines. Neural Networks for Signal Processing, 511–520 (1997)
11. Murzin, A., Brenner, S., Hubbard, T., Chothia, C.: SCOP: A Structural Classification of Proteins Database for Investigation of Sequences and Structures. Journal of Molecular Biology **247**, 536–540 (1995)
12. NEC Labs America. http://ml.nec-labs.com/download/data/svm+/mnist.priviledged
13. Ortiz, A., Strauss, C., Olmea, O.: MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison. Protein Science **11**, 2606–2621 (2002)
14. Pechyony, D., Izmailov, R., Vashist, A., Vapnik, V.: SMO-style Algorithms for Learning Using Privileged Information. In: 2010 International Conference on Data Mining, pp. 235–241 (2010)
15. Tsybakov, A.: Optimal Aggregation of Classifiers in Statistical Learning. Annals of Statistcs **31**, 135–166 (2004)
16. Schölkopf, B., Herbrich, R., Smola, A.J.: A Generalized Representer Theorem. In: Helmbold, D.P., Williamson, B. (eds.) COLT 2001 and EuroCOLT 2001. LNCS (LNAI), vol. 2111, p. 416. Springer, Heidelberg (2001)
17. Steinwart, I., Scovel, C. When do support machines learn fast?. In: 16th International Symposium on Mathematical Theory of Networks and Systems (2004)
18. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)
19. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
20. Vapnik, V.: Estimation of Dependencies Based on Empirical. Data (2nd Edition). Springer, New York (2006)
21. Vapnik, V., Chervonenkis, A.: Theory of Pattern Recognition (in Russian). Nauka, Moscow (1974). German translation: Wapnik W., Tscherwonenkis, A.: Theorie des Zeichenerkennung. Akademie-Verlag, Berlin (1974)

22. Vapnik, V., Vashist, A.: A New Learning Paradigm: Learning Using Privileged Information. Neural Networks **22**, 546–557 (2009)
23. Vapnik, V., Izmailov, R.: Statistical inference problems and their rigorous solutions. In: Gammerman, A., Vovk, V., Papadopoulos, H. (eds.) Statistical Learning and Data Sciences. SLDS 2015, LNCS (LNAI), vol. 9047, pp. 33–71. Springer-Verlag, London (2015)
24. Vapnik, V., Braga, I., Izmailov, R.: A Constructive Setting for the Problem of Density Ratio Estimation. In: SIAM International Conference on Data Mining, pp. 434–442 (2014)