Regularizing Pattern Recognition with Conditional Probability Estimates

1st Thomas Vacek
This work was completed independently and unaffiliated
Minneapolis, MN USA
twvacek@gmail.com

***Abstract*—Recent contributions in non-parametric statistical pattern recognition have investigated augmenting the task with information about the conditional probability distribution $\mathbf{P(Y|X)}$ away from the $\mathbf{0.5}$ level set, i.e. the decision boundary. Many hypothesis spaces satisfy generous smoothness criteria, so the behavior of a function away from the decision boundary can serve as a regularizer for its behavior at the decision boundary. This paper proposes a paradigm to capture observable information about the conditional distribution and describe a learning formulation that can take advantage of it. Finally, it investigates why conditional probability can be an effective regularizer for inseparable pattern recognition problems.**

I. Introduction

The binary pattern recognition task is to approximate the 0.5 level set of the conditional probability function $P(Y = 1|X) \coloneqq \eta(X)$, where P is some distribution that jointly generates labels Y in ± 1 and observations X in $\mathbb{R}^n$. It is assumed that labels are difficult or costly to observe, while observations are commonplace, giving rise to value for a machine that can accurately predict the label for a given example.

The methods and analysis proposed here pertain generally to any pattern recognition formulation which satisfies the following axioms: (1) A hypothesis h (a.k.a. discriminant, index, or network in various communities) is to be chosen from a large class $\mathcal{H}$, (2) so that some level set of h approximates the $.5$-level set of η, (3) where all elements of $\mathcal{H}$ satisfy some smoothness property.[1] Formulations which satisfy these axioms are popular and widespread, such as kernel methods, support vector methods, generalized linear models, and artificial neural networks. While this paper implements and evaluates formulations using kernel and support vector methods, many of the results are theoretical and thus likely to apply more generally.

In the development of classical non-parametric statistical pattern recognition, the behavior of the conditional distribution away from the decision boundary ($|\eta - .5| \gg 0$) has not received much attention. VC dimension [1], a generalization of linear dimension, is defined in terms of the decision rule induced from the hypothesis (a characteristic function of the hypothesis), so that the behavior of the hypothesis away from the zero level set is inconsequential to VC dimension, except to the extent that it is determined by smoothness properties. By contrast, this paper explores whether constraints or penalties on the behavior of a hypothesis away from the decision boundary will improve our ability to estimate the behavior of the decision boundary, assuming the hypothesis is appropriately smooth. The term *conditional probability regularization* (CPR) is coined for this technique. The CPR method can be understood intuitively as to squish the dispersion of the chosen hypothesis $h(X)$ when graphed against the conditional probability $\eta(X)$, as illustrated in Fig.I. Typical pattern recognition loss functions, such as hinge loss or binomial log likelihood attempt to reduce the density of the hypothesis h in the neighborhood of 0, making a scatterplot as steep as possible in the vicinity of the decision boundary ($\eta = .5$), up to the allowable regularization. In contradistinction, the proposed regularization trades off the approximate slope of the scatterplot with the dispersion, as best we can estimate these quantities.

As a fundamental matter, if $|\eta - .5|$ is bounded far away from zero, CPR methods are not useful. In these circumstances, the location of the decision boundary is not as important [18], so careful placement of it is not useful. Thus, theory suggests that CPR methods are only effective for moderately or highly inseparable tasks, when the loss of the best models is high.

The author argues that the following principles are desirable for any CPR formulation:

1) Empirical data required by the technique should be possible to obtain for reasonable pattern recognition tasks, and the cost of collecting it should be not much worse than the cost of collecting ground-truth class labels.
2) The regularization should be statistically efficient; that is, there should be some provable mechanism to show that the CPR allows one to choose a better hypothesis for a given sample than without it under practically attainable assumptions. Additionally, if the assumptions do not hold, the formulation should fail gracefully and produce a model not much worse than an efficient formulation which does not use CPR.

This paper addresses these *desiderata* in the following ways: First, empirical data are assumed to be a noisy, unknown

[1] This paper does not attempt to study specific continuity or smoothness properties; the proposed formulations used all use linear or RBF kernels, which are analytic functions. The authors believe that the method would apply to piecewise smooth functions and likely to Hölder- and Lipschitz-continuous functions as well.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

increasing transformation of the conditional probability η for each example in the training set. Note that an unknown decreasing transformation can be converted into an unknown increasing transformation via multiplication by -1. The method also allows a different transformation of η for each label class.

While true values of η are rarely observable, there are common phenomena that can be interpreted as noisy transformations, especially in tasks which relax the regression of an unknown function into predicting the sign of that function. Observable values of the regression can be interpreted as an unknown transformation of η. A canonical example is the pattern recognition task, which is to find the sign of the unknown function $\eta - .5$

For example, consider a patient who survives 30 years after a serious diagnosis compared to a patient who survives five years. If the goal is to predict five-year survival, patients essentially like the former have a higher probability of surviving than patients like the latter. Thus, the length of survival after diagnosis can be interpreted as a noisy increasing transformation of a the conditional probability of surviving five years.

For the second *desiderata*, this paper justifies the CPR intuition using well-known results from fast-converging excess risk bound literature. It shows how learning formulations which satisfy the 3 axioms in the introduction can be coerced into satisfying the nontrivial requirements for these bounds. Coercion is accomplished by the addition of constraints based on empirical conditional probability data. Practically, constraints are converted to parametric penalties. When fully relaxed, the penalties vanish leaving the original learning formulation. This paper chooses penalty parameters empirically via validation.

A basic question will naturally arises from the problem setting just described: Why would one want to find the sign of an unknown function instead of that function directly? For example, it would be possible to predict a patient's survival time, either directly or as a survival curve. The author submits three reasons: First, it is assumed that the regression task requires a more complex function to imitate it, so there is a higher risk of underfitting or overfitting if one only seeks the sign to begin with. Second, a model may seek agreement with human judgments, which routinely distill complex, continuous-valued phenomena into a binary decisions, such as whether a student passed a test or whether a party was liable in a legal case. In tasks like these, the underlying preferences and reasons may be too complicated and costly to observe directly, or even apparently inconsistent, even though human decision-makers have no trouble describing subjective confidence in their decision. Finally, in applications where the regression task is used as one input in a sequence of decisions, its sign may be all that is needed.

This paper proposes two CPR methods, provides some theoretical justification in terms of excess risk bound literature (though the author is aware of many additional results) and evaluates them against a baseline non-CPR pattern recognition method and, for some experiments, against a decision rule induced from a regression on empirical conditional probability. The evaluation also studies the estimation error that arises from having to choose CPR penalty parameters empirically. The methods proposed here are based on kernel and support vector methods. This combination gives rise to smooth hypothesis spaces and convex inference formulations. The latter property is important in evaluating a regularization strategy, as nonconvex formulations present competing explanations to differences in performance. The codes used here are for research and are not optimized for large datasets.

II. Background

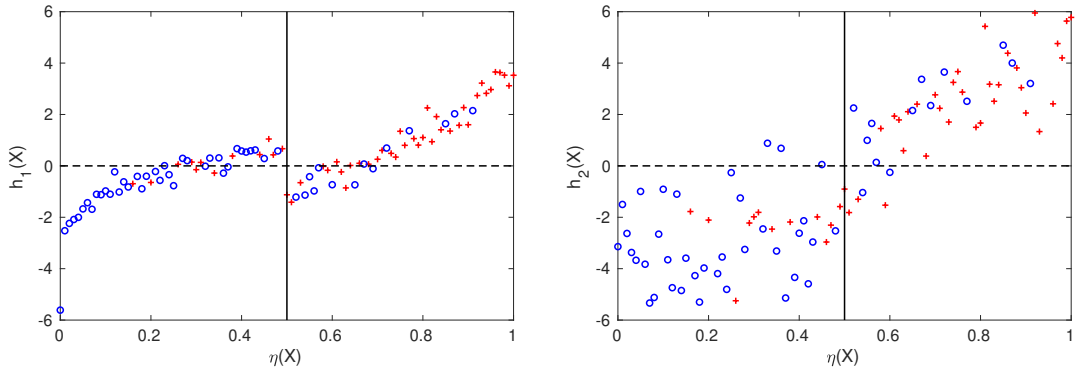
Vapnik and Vashist [2] introduced the idea of privileged information, which is a general term for information about an observation X which is not readily observable but which can be assumed in finite quantities. There are no further formal assumptions. It may be high dimensional. Statistical learning formulations making use of privileged information (called Learning Using Privileged Information or LUPI) should have particular assumptions as how the privileged information can be leveraged. In this terminology, privileged information is distinguished from decision information, which is the observation X.

A common assumption is the teacher assumption: there is some admissible function under which the privileged information is more predictive than the the desired task. The teacher function applied to an observation's privileged information gives rise to constraints that the hypothesis for the desired task (decision) hypothesis should satisfy. For example, in the medical field, tumor diagnosis is a sequential decision proceeding from imaging to biopsy and ultimately surgery, reflecting a progression of increasing cost, invasiveness, and certainty. It is useful to distinguish three quantities: $P(Y|X)$, $P(Y|X^{obs})$, and $P(Y|X^{priv})$. In the example, the first value would be the probability that a tumor is malignant, taking all possibly-observable information into account. The second might be the probability of malignancy based only on imaging. The last might be the probability of malignancy given the knowledge gained from a more intensive investigation, such as a biopsy. A decision rule for a biopsy, along with the results of a test, could be a teacher for detecting malignancy from images. Generally speaking, the three conditional probabilities are all different functions, and most papers in the literature differ in how to express a teacher assumption in terms of these quantities. This paper is distinguishable because it assumes direct empirical access to $P(Y|X^{obs})$,[2] as opposed to indirect access to $P(Y|X^{priv})$ via some admissible function. Although empirical conditional probability information is one of many possible kinds of privileged information, this paper prefers to use the former term for its specificity.

The paper which proposed the LUPI paradigm [2] also proposed a learning formulation called SVM+ which seeks an admissible function of the privileged information to explain

[2]In some cases, empirical conditional information is in the form of $P(Y|X)$. This paper regards deviations between it and $P(Y|X^{obs})$ as noise.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE



(a) (b)

Fig. 1. Scatterplots comparing hypothetical examples of conditional probability regularization. Conditional probability $P(Y = 1|X) \coloneqq \eta(X)$ is plotted against the behavior of a hypothetical discriminant function h on a synthetic sample. The scatterplot markers $+$ and $\circ$ depict the two classes of the pattern recognition problem. A good discriminant minimizes the density in the diagonal quadrants; an optimal discriminant (Bayes rule) has none. A task with low loss would have low density in the center of the scatterplot. In the first instance, h_1 has low dispersion with respect to the true conditional probability, but is is inversely related to η in the vicinity of $\eta = 0.5$, so has undesirable high density in the diagonal quadrants. In the second example, h_2 has higher dispersion with respect to η, but better achieves the desired low density of the diagonal quadrants.

the losses of a good pattern recognition hypothesis. This is not explicitly a teacher assumption because the chosen function need not perform well on the desired task, although the possibility is not excluded. However, the loss of a good pattern recognition hypothesis should be related to conditional probability, so the teacher assumption is arguably implicit here.

Pechyony and Vapnik [3] explained a mechanism to prove that SVM+ can have convergence of empirical risk to expectation at a rate like $\frac{1}{n}$, and provided an example of a distribution that is provably slow ($\frac{1}{\sqrt{n}}$) for a learner without privileged information and provably fast with SVM+. However, neither this nor any other formulation to the author's knowledge prove fast convergence results based on criteria that can be verified from data, except for the results in this paper.

Wang and Ji [9] propose a formulation (LIR) based on the teacher assumption. They assume a principle that the student should not be better than the teacher on any training point; that is, the student's loss should not be less than the teacher's. A similar condition could be expressed in terms of conditional probability.

Lopez-Paz, Schölkopf, Bottou, and Vapnik [5] relate model distillation (using a complex teacher model to create pseudo-labeled data for a simpler model) to the teacher assumption, where the simpler model is required to emulate the conditional probability of the teacher model. Suppose that L_n^{targ} is an empirical risk functional for the target task and that L_n^{teach} is an empirical risk functional for the deviation between the predicted conditional probability of an example under the complex and simple models. Then the formulation can be written as:

$$\min_{h \in \mathcal{H}} L_n^{targ}(h) + C L_n^{teach}(h)$$

The paper suggests annealing the extreme values of of the teacher conditional probability, similar to the idea that a student should not be more certain than the teacher.

Wang et al. [7] finds a favorable partition of the privileged space using a linear latent variable model and seeks to predict the partition in the decision space, in the hopes that a partition-dependent variables can lower loss compared to a global model. This is a teacher approach, but the feedback between the teacher and student is through the partition rather than conditional probability or loss.

Other formulations use the privileged information as a guide, but do not require finding an admissible function that is predictive of the desired task. Wang and Ji [9] assumes that propose a formulation (RPR) that interprets the privileged information as allowable components of the loss on the training sample.

Lapin, Hein, and Schiele [4] propose to use empirical information as conditional probability estimates, similar to this paper but not allowing for transformation invariants. They find that the SVM+ method is reducible to a weighted SVM approach (though finding equivalent weights is not trivial) and suggest that the best weights are conditional probability estimates. The method encourages a learner to prioritize performance on the easy examples over the hard examples. However, this approach requires accurate conditional probability estimates, excluding some sources of this information considered here.

Vapnik and Izmailov [10] have recently proposed a Learning Using Statistical Invariants framework. This approach attempts to directly solve the ill-posed estimation problem for the conditional probability function η. Since an empirical estimate of η is found in some parametric form, additional *a priori* constraints can be placed on it, which are called statistical invariants. The method proposed here differs in that the hypothesis selected is constrained by empirical estimates of

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

η, rather than *a priori* properties, and the hypothesis here is not an approximation of η.

There some papers that study empirical conditional probability or privileged information, but which use hypotheses that do not satisfy our smoothness assumptions. Some of these use decision tree methods [6, 11], while Fouad et al. [8] considers using privileged information to adapt a distance function for prototype learning (methods such as K-Nearest Neighbors).

III. Theoretical considerations

The first risk bounds to be proved [1] are called *uniform bounds* because they bound the deviation between empirical risk and the risk in expectation uniformly and simultaneously for all admissible functions in the space under consideration. That is, with high probability $\sup_{h\in\mathcal{H}} L_n(h) - L(h) \leq B$, in other words, the empirical risk (loss on a finite sample) will not exceed the true loss (supposing one had access to the true distribution). This approach applies quite generally to any bounded loss function, provided that VC-dimension is controlled.

Another line of analysis is to consider the *excess risk* of a hypothesis chosen according to the empirical risk minimization (ERM) procedure: Let h_n minimize L_n (empirical risk) and h' minimize L (loss expectation). An excess risk bound has the form $L(h_n) - L(h') \leq B$. This bound can be tighter than the uniform bound; the idea is that the 'bad' hypotheses that cause the uniform bound to be large might have high loss and would be excluded by the ERM procedure. The perfection of this idea is known as the 'peeling' procedure. Namely, if one can control the variance of the excess risk in terms of its expectation uniformly over $h \in \mathcal{H}$, such as $\mathrm{Var}[l(h) - l(h')] \leq D\,\mathrm{E}[l(h) - l(h')]$, then excess risk bound converges like $\frac{1}{n}$.[3] Fast-converging excess risk bounds, as presented above, apply to theoretical 0/1 loss, while practical learning formulations almost always use a convex relaxation such as hinge loss or negative binomial log likelihood. In some cases, the fast convergence argument can be extended to the the relaxed loss [13]. To the author's knowledge, the typical means of extending fast convergence proofs to convex loss cannot be applied to the formulations proposed here. Nevertheless, other things equal, a formulation which has provable fast convergence properties in its unrelaxed form is superior to one which does not. It is possible that advances in computing will make it tractable to optimize for 0/1 loss.

The key mechanism to do a theoretical analysis on the proposed problem setting is the following: If a total order can be induced on a distribution and every hypothesis in a space has 0/1 loss which respects that order, then the variance condition is trivially satisfied. For example, suppose we seek to find the best hypothesis over a discrete distribution of three points with a total ordering defined by $(y_1, x_1) \succ (y_2, x_2) \succ (y_3, x_3)$. Then the variance condition is satisfied ($D = 1$) for 0/1 loss if for every $h \in \mathcal{H}$, $l(h(x_1), y_1) \geq l(h(x_2), y_2) \geq l(h(x_3), y_3)$. The proof of this is trivial. An extension of this principle is that the variance condition can be satisfied by ordered hypotheses, as in for every $h \in \mathcal{H}$, $h(x_1) \geq h(x_2) \geq h(x_3)$, with some additional assumptions required.

The author proposes two formulations based on these mechanisms: First, observe that any hypothesis h can be coerced into having ordered loss by the addition of some non-negative correcting function: $l(h) + \phi(h, \omega)$ where ω is a parameter.[4] The realization of this idea is called *LO-SVM*. Second, empirical risk minimization can be performed over a filtered hypothesis space, where every admissible hypothesis satisfies a relaxed order condition with high probability:

$$\min_{L^{order}(h)\leq\gamma} L^{targ}(h)$$

In our case, L^{targ} is 0/1 loss, and L^{order} is an ordinal regression formulation. (Use of ordinal regression here allows the formulation to enjoy invariance to increasing transformations in the empirical conditional probability information.) For L^{order}, a useable formulation based on pairwise loss was analyzed by Clémençon et al. [14] using the theory of U-statistics proposed by Hoeffding [15]. This paper proposes a modified formulation called *Asymmetric Pairwise Loss* (APL). The formulation arising from this is called *GO-SVM*.[5]

To turn these into a useful formulation, the following steps are needed.[6]

- The 'peeling' argument for fast convergence can be slightly relaxed to allow for a less than perfect total ordering.
- The requirement of ordered loss is relaxed to ordered loss in each class. This expands the kind of problems where the method can be applied and can be shown to not substantially affect rate of convergence.

The analysis proposed here might be called *risk bound engineering* in that we can take well-known proofs and combine a number of them using the union bound. Since risk bounds are inherently conservative, using the union bound only exacerbates the problem, and and many of the proofs are not at all tight. Still, the exercise is useful to understand the implications and tradeoffs of the methods.

IV. LO-SVM

The LO-SVM formulation[7] takes advantage of the fact that every hypothesis in a space can be coerced to have loss which satisfies a desired order. Assume data are generated according

[3]Relaxed versions of this condition are also possible. The criterion is most commonly presented in terms of the Mammen-Tsybakov noise conditions [12], which further assume $h^* \in \mathcal{H}$. This latter assumption is a tool to prove the variance criterion, not a requirement of the peeling proof.

[4]This is the same principle used by Vapnik's SVM+ [2]. In this case, ω defines a threshold of the total ordering, below which there is loss and above which there is not.

[5]The names are intended to be mnemonic for *loss order* and *global order* respectively. The GO-SVM requires that every point respect the ordering whereas the ordering constraints in LO-SVM are defined by the points which have loss.

[6]We omit proofs here because they are tedious and not germane to the intended audience. Interested readers can find them in [16].

[7]The LO-SVM formulation can be regarded as a specialization of the SVM+ algorithm in [2] with the correcting function space replaced by $\mathcal{M}$, the class of 1-dimensional increasing functions.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

to a a distribution P which generates points $(X, Y, R) \in \mathbb{R}^d \times \pm 1 \times \mathbb{R}$. The last argument is the conditional probability proxy, encoded as a real number.

A classical VC-analysis is straightforward, so it will be outlined. Let $\mathbb{1}_X$ be a characteristic function that is 1 on the support of X; let E be statistical expectation; let E_n be empirical expectation on a random sample of size n; let l^{01} be standard 0/1 loss; finally, we use the shorthand $L^{01} \coloneqq \mathrm{E}[l^{01}]$. Finally, let h_n be a hypothesis (and similarly for other parameters) generated by empirical risk minimization over a sample of size n. Consider a correcting function of the form:

$$c(x, y, r; h, \omega) = \mathbb{1}_{r \leq \omega} - l^{01}(h(x), y)$$

where ω is a parameter in $\mathbb{R}$. Then it is trivial to write a formulation in expectation:

$$\min_{h,\omega} \mathrm{E}[l^{01}(h) + c(h, \omega)] = \mathrm{E}[\mathbb{1}_{r \leq \omega}]$$
$$\text{s.t. } c(x, y, r; h, \omega) \geq 0$$

Note that the quantity to be minimized is an upper bound on the loss of $L^{01}(h^*)$, and it satisfies the ordered loss condition and a has a VC dimension of 1 (regardless of the hypothesis space being used). Thus, fast excess risk bound theorems hold here, and the empirical formulation will converge at a $\frac{1}{n}$ rate to its minimum admissible value in expectation, provided that uniform non-negativity of the correcting function can be guaranteed. Uniform non-negativity of $c(h_n, \omega_n)$ cannot be guaranteed by empirical methods; however, the probability of negativity can be bounded which then yields a useful risk bound.

A result of Bartlett et al. [17, Corollary 3.7] gives a bound that an independent constant C exists so that with probability at least $1 - \delta$, uniformly for all $(h, \omega) \in \mathcal{H} \times \mathbb{R}$ such that $E_n[\mathbb{1}_{c(h,\omega)<0}] = 0$ the following holds: $\mathrm{E}[\mathbb{1}_{c(h,\omega)<0}] \leq C\frac{(V+1)(\log \frac{n}{V+1})+\log \frac{1}{\delta}}{n}$. Let $\mathring{h}, \mathring{\omega}$ minimize $E[f]$ over all hypotheses for which $E_n[\mathbb{1}_{c(h,\omega)<0}] = 0$ (note that this is defined on the training set).[8] A standard variance bound result of Boucheron et al. [18, Section 5.3.4] is that there is an independent constant C such that with probability at least 1-α,

$$\begin{aligned}\sup(\mathrm{E}[f(h_n, \omega_n)] - \underset{n}{\mathrm{E}}[f(h_n, \omega_n)],& \\ \mathrm{E}[f(h_n, \omega_n)] - \mathrm{E}[f(\mathring{h}, \mathring{\omega})])& \\ \leq C\frac{\log n + (\log n + 4)\log \frac{1}{\alpha}}{n}&\end{aligned}$$

The two bounds hold simultaneously with probability at least $1 - \alpha - \delta$, and adding the right hand sides gives us a fast $\frac{1}{n}$ upper bound on the convergence of h_n to $\mathring{h}$, as defined here. Unfortunately, $c(\mathring{h}, \mathring{\omega})$ might be quite large, so there is no reason to believe h_n converges to the minimizer of $L^{01}(h)$ without further assumptions. To satisfy the second axiom of CPR formulations, that they should fail gracefully, penalized negativity of c can be permitted; that is, certain examples are allowed to have loss. This reduces the speed of convergence of the objective and loosens the bound on the nonnegativity of c, but has the benefit of tightening the upper bound.

A benefit of this method is that the assumption that CPR information provides a total ordering of the training examples can be relaxed to an assumption that the CPR information gives two total orderings, one for each class. The analysis described above has to be adapted. This is not trivial; one has to avoid a pathology where a hypothesis space contains two hypotheses with similar loss but where the loss in each class is much different. However, this requirement can be empirically enforced as well, and then added to the union bound.

This method was implemented using by relaxing 0/1 loss to hinge loss. The ordering constraints are assumed to be per-class. The loss balance constraint is not explicitly enforced because the optimal models without the constraint have reasonable balance. The non-negativity constraint on c is relaxed via a penalization parameter C. As in ν-SVM, the hypothesis space complexity is traded off with loss via a parameter ν. The formulation is:

$$\begin{aligned}\min\ & \frac{1}{2}w^T w - \nu\rho + \frac{C}{n}\sum_i \zeta_i + \frac{1}{n}\sum_i \xi_i \\ \text{s.t. } & \forall i, y_i(w \cdot x_i + b) \geq 1 - \zeta_i - \xi_i \\ & \xi \geq 0 \\ & \zeta \geq 0 \\ & \forall (i, j) \in \mathcal{P}, \zeta_i \leq \zeta_j\end{aligned}$$

The preference set $\mathcal{P}$ is defines chained constraints based on a (per-class) total ordering, as in $\zeta_1 \leq \zeta_2 \leq \zeta_3 \leq \ldots$. Relaxations of the ordering are computationally straightforward, although a statistical analysis is more difficult. If $C \geq 1$, the method is the same as ν-SVM [21]. The formulation can be written using two cost parameters, but is apparently sensitive. The author found that the ν-formulation found better models for all tasks in the evaluation.

The Representer Theorem [19] holds for this formulation, and the method is implemented in the dual form, allowing kernel learning. Model inference requires solving a convex quadratic program with n (sample size) variables and a multiple of n constraints.

V. GO-SVM

The GO-SVM formulation applies two different kinds of regularized loss to the same hypothesis, namely L_n^{01} and L_n^{teach}. In principle, any regression formulation could be used for L_n^{teach}, but there are a some considerations:

- A method which is insensitive to increasing transformations of empirical conditional probability data requires a special formulation. Pairwise loss and the variant defined here are the only possibilities known to the author.
- Both L_n^{01} and L_n^{teach} should be regularized; however, the regularizations of each loss cannot conflict. A regularization that makes use of the norm of the hypothesis cannot

[8] A necessary step is to establish the VC dimension of the class $\mathcal{C}(\mathcal{H}) = \{\mathbb{1}_{c(h,\omega)<0} : h \in \mathcal{H}, \omega \in \mathbb{R}\}$. A similar result is proved for pseudo-dimension by Pechyony and Vapnik [3, Appendix B] which can be specialized to VC dimension. The VC-dimension of $\mathcal{C}(\mathcal{H})$ is bounded by $V + 1$.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

coexist with another. To the author's knowledge, the only loss formulations that allow independent regularization are ones in which a capacity parameter is an explicit optimization variable, such as ν-SVM methods [21].

This section defines a new ordinal regression loss called Asymmetric Pairwise Loss (L^{apl}). It then discusses how regularized formulations of L^{apl} can be combined with a regularized L_n^{01} to make a learning formulation.

Asymmetric Pairwise Loss (L^{apl}) is introduced here as the unrelaxed (non-convex) prototype of the regularized empirical ordinal regression formulations defined by Shashua and Levin [20]. They proposed two formulations, both of which can be viewed as regularized convex relaxations of L^{apl}. Unlike the common ordinal regression definition, this author does not assume that any point in the support of the target distribution has measureable mass; this interpretation of ordinal regression as scale-insensitive regression was given in Clémençon, Lugosi, and Vayatis [14]. The core of APL is an indicator function that indicates loss if a threshold t splits a prediction $f(x)$ from its target y.

Definition 1 (Order Contradiction Indicator)**.**

$$l^{oci}(f(x), y; t) := \begin{cases} 1 & f(x) > t, y < t \\ 1 & f(x) < t, y > t \\ 0 & \textit{otherwise}. \end{cases}$$

With this definition in hand, we can define Asymmetric Pairwise Loss:

Definition 2 (Asymmetric Pairwise Loss)**.**

$$L^{apl}(f) = \mathop{\mathrm{E}}_{X',Y'}\left[\mathop{\mathrm{E}}_{X,Y}\left[l^{oci}(f(X), Y; Y'\right]\right]$$

The definition is expressed as a double expectation for clarity; however, X' is trivial and could be dropped from the outer expectation. Intuitively, error is the expectation over predictions and labels of the mass of points which induce order contradictions. The analysis by Clémençon, Lugosi, and Vayatis [14] can be extended to this approach, and APL satisfies a strong risk bound under heteroscedastic noise provided that it is mean zero with bounded variance at each point. (Pairwise loss requires assuming symmetric noise according to their analysis.) Moreover they suggest that the variance properties of the U-statistic may be more favorable than their analysis can prove.

The fixed-margin formulation of Shashua and Levin [20] assigns each training example a distinct interval of some minimum size (margin) on a real number line that is consistent with its target label. Loss is the magnitude of the displacement an example from its interval, and it is empirically minimized by optimizing the intervals and the hypothesis that projects each example onto the number line. The displacements are relative to the minimum interval size, so loss is proportional to the mass of contradicting points if Y is distributed uniformly. For other distributions the displacements can be interpreted as an upper bound to the mass of contradicting points with appropriate assumptions.

A risk bound for GO-SVM is straightforward using the techniques described for LO-SVM when the ordinal component of the problem has zero loss, which is not a plausible assumption. This can be relaxed, but the resulting bound is not fast for 0/1 loss below the level of relaxed ordinal loss.[9] In the author's opinion, the variance of hypotheses in the classification task is much more controlled than the analysis can show. Unlike LO-SVM, this method directly minimizes the pattern recognition loss instead of an upper bound.

The GO-SVM formulation uses a linear hypothesis w (subject to the kernel trick) as in SVM. Loss and capacity control are traded off between hinge loss (relaxed 0/1 loss) and ordinal loss via user-selectable parameters. Capacity control in both SVM and the ordinal regression formulations is attained by the relationship between the squared norm of the predictor w and the size of the margin. However, w serves a two-fold role in this formulation; therefore implementing different capacities for the two learning objectives requires explicit values for the margins. This formulation extends ν-SVM formulations [21] so that the usual tradeoff between loss and capacity is preserved.

The formulation is

$$\begin{aligned} \min_{\substack{\xi \geq 0 \\ w,b,g,\xi \\ \zeta,\rho_b,\rho_o}} \; & \frac{1}{2} w^T w + \alpha \left(-\nu_b \rho_b + \frac{1}{n} \sum_{i=1}^{n} \xi_i \right) \\ & + (1-\alpha) \left(-\nu_o \rho_o + \frac{1}{n^*} \sum_{i=1}^{n} |\zeta_i| \right) \\ \text{s.t. } & \forall i,\; y_i (w \cdot x_i + b) \geq \rho_b - \xi_i \\ & \forall i,\; g_i + \frac{\rho_o}{2} \leq w \cdot x_i + \zeta_i \leq g_{i+1} - \frac{\rho_o}{2} \end{aligned}$$

Hinge loss is implemented in the first and third lines, while ordinal loss is implemented in the second and fourth lines. Variable g is the vector of interval boundaries in natural order, with one example per interval if no ties and assuming no empty intervals. Extending the formulation to support two within-class orderings can be accomplished by a straightforward adaptation. Variable w is the linear hypothesis and b is a classification bias term. Constant n^* is defined to control the feasible range of ν_o. It is ${}^{n^2-n}/_2$ if there are not ties in the ordering.

Parameter ν_b controls the VC-dimension of the 0/1 loss class, ν_o controls the VC-dimension of the regression task hypotheses. Finally, parameter α trades off the loss on the target decision task with the loss on the CPR regularizer regression task. In principle, within-class orderings require enforcement of loss balance. Loss balance is not enforced; however, the author has never observed the unconstrained optimum to have unreasonable loss balance, and the value of the unconstrained minimizer can be thought of as a post-hoc constraint.

Like ν-SVM [21], the optimization problem can be characterized in terms of ν_b and ν_o and training data.

[9] This bound could be useful for tasks with high loss on the classification task, but low loss on within-class ordinal tasks.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

It can be proved that the problem is primal and dual feasible for $\nu_o \in [0,1]$, $\alpha \in [0,1]$, and $\nu_b \in [0, 2\min(\#\text{ positive examples}, \#\text{ negative examples})/n$; and primal unbounded/dual infeasible otherwise. The Representer Theorem [19] holds for GO-SVM, so the solution can be expressed in terms of the dual variables and kernels can be used.

VI. Evaluation

The goal of evaluation is to prove that the conditional probability regularizer allows faster convergence (empirical risk to its expectation) than a learning formulation which considers only the labels. Since both proposed CPR formulations are an extension of standard SVM, it is a logical baseline. Moreover, because the LO-SVM and GO-SVM hypothesis spaces are restrictions of the full hypothesis space in SVM, they cannot outperform SVM by virtue of a richer hypothesis space; SVM will always attain the lowest loss on the training sample. Empirical performance improvements over the baseline can only be explained by better convergence. Finally, evaluation is not intended to be a statement about the fitness of the hypothesis spaces for the learning task, but only about the ability of the learner to select the best element. The author believes the results here would carry over to hypothesis spaces that have evolved to support specialized tasks.

The baseline solver is a ν-SVM formulation that is coded using Matlab's quadprog interior-point convex solver, as are all formulations used here. All optimization tasks used in these experiments operate on the same kernel matrix and solve to the same numerical tolerance.

The experimental setup is to hold out a testing set and sample remaining examples for 20 random realizations of training and validation sets. The validation sets were used for model selection, and results are reported on the test set, which is used for all experiments. Testing sets contained at least 1800 examples. Experiments use two validation sizes. The first validation set is the same size as the training set and intended to emulate cross validation but save computing time. The second is much larger and intended to show how the methods would perform if model selection were optimal. These results are reported as 'holdout' and 'extended' experiments, respectively.

The ν formulations have a fixed, auto-scaling parameter ν, and we use structural risk minimization to choose from a fixed set of parameters $\nu = [.1, .2, \ldots, .9, .95, .995]$. The LO-SVM formulation performs grid search model selection using these ν parameters and a C parameter set of $[.5.7.8.9]$. We found that the LO-SVM problem was often infeasible for values of C, but this paper did not attempt to study its feasibility.

The rbf kernel width (where used) is chosen from the $[.1, .25, .5]$-quantiles of the pairwise distance of training points. The kernel parameter was chosen by a hold-out validation on the SVM experiment and re-used in the other formulations to cut down the size of the model search. The α parameter in the GO-SVM method was chosen from $[.1, .25, .5]$.

A. Synthetic data

A synthetic dataset is designed to illustrate the properties of the method in terms of underlying statistical properties of data. The dataset is defined using a distribution to sample $P(Y = 1)$ and then to generate X and Y such that the sampled value represents $\eta \coloneqq P(Y = 1|X)$. The signal is obfuscated by rotating with high dimensional noise via a random orthogonal matrix Q.

A dataset is generated by first defining a distribution $\hat{\eta}$ with support in $[0, 1]$ and a random orthogonal matrix Q with dimension d. Individual examples are generated with the following steps:

1) Sample a base value $T \sim \hat{\eta}$.
2) Assign a class label 1 according to the probability T.
3) Sample a uniform random vector U of size $d - 1$ and let $X = Q\,[T; U]$.
4) Create empirical conditional probability labels by adding noise to the base value T, as $R \sim T + \mathcal{N}(0, \sigma)$.

This procedure implies that $\eta(X) = mX$ for some m.

These experiments use a simple $\hat{\eta}$ distribution defined by parameter θ. This distribution has uniform density on $[0, 1]$, except for a tooth in the interval $[\theta, 1 - \theta]$ with no support. When θ is small, the task is easy and the variance condition for fast convergence holds more strongly. As θ approaches .5, the task is difficult, as a significant fraction of labels are essentially random. These distributions are symmetric around .5, so they generate class-balanced problems. It is straightforward to compute the Bayes classifier error rate (which assumes η is known) in terms of θ.

While ν-SVM is the baseline for all experiments in this paper, the synthetic experiments include a conditional probability regression approach. This formulation can be viewed as the L^{teach} component of the proposed GO-SVM formulation. However, the CPR formulations assume that empirical conditional probability values are not comparable across classes. The ordinal regression baseline does not use stronger assumptions; rather, the two preferences orderings are glued together by placing the negative class ordering before the positive class ordering. After the optimal regression function is found, a decision boundary is chosen based on training data.

The synthetic evaluation considers a very small sample of size 30 and dimension of size 10 for values of θ at 0.3, 0.45, and 0.5. For each of these, we consider empirical conditional probability data with and without noise ($\sigma = .1$). As θ increases, the loss of the Bayes optimal decision increases, and the task becomes more difficult. The experiment also includes evaluation of increasing conditional probability noise for $\theta = .45$.

B. Common datasets

The evaluation also includes a some datasets from related papers, as well as a few UCI datasets that can be adapted to the CPR data assumptions. An SVM+ implementation was also evaluated, using the empirical conditional probability

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I

RESULTS (MEAN ERROR RATE AND STANDARD DEVIATION OVER 20 RANDOM EXPERIMENTS FOR SYNTHETIC DATASET FOR VARIOUS DISTRIBUTIONS. WINNERS ARE REPORTED INDEPENDENTLY FOR HOLDOUT AND EXTENDED VALIDATION EXPERIMENTS. ALL EXPERIMENTS HAVE SAMPLE SIZE 30, DIMENSION 10

θ, σ	ν-SVM holdout	ν-SVM extended	GO-SVM holdout	GO-SVM extended	LO-SVM holdout	LO-SVM extended	OR-SVM holdout	OR-SVM extended	Bayes optimal
$0.3, 0.0$	.188 (.031)	.161 (.014)	.187 (.031)	*.150 (.000)*	**.168 (.028)**	.151 (.003)	**.168 (.017)**	.161 (.011)	.15
$0.3, 0.1$	.179 (.039)	.165 (.019)	.171 (.028)	*.152 (.002)*	**.168 (.031)**	.155 (.006)	.189 (.029)	.167 (.011)	.15
$0.45, 0.0$	.291 (.049)	.269 (.042)	**.272 (.031)**	*.232 (.007)*	.279 (.040)	.260 (.033)	.282 (.041)	.265 (.024)	.225
$0.45, 0.1$	.322 (.061)	.299 (.041)	**.259 (030)**	*.230 (.002)*	.278 (.025)	.254 (.024)	.277 (.039)	.262 (.026)	.225
$0.45, 0.2$	.295 (.055)	.277 (.052)	**.256 (.024)**	*.228 (.004)*	.279 (.043)	.256 (.028)	.274 (.039)	.255 (.032)	.225
$0.45, 0.3$	.308 (.031)	.271 (.028)	**.276 (.035)**	*.243 (.011)*	.291 (.031)	.262 (.022)	.285 (.031)	.261 (.016)	.225
$0.45, 0.4$	.288 (.033)	.268 (.025)	**.285 (.036)**	*.250 (0.18)*	.294 (.028)	.265 (.024)	.295 (.054)	.270 (024)	.225
$0.45, 0.5$	.311 (.067)	.279 (.036)	**.304 (.065)**	*.255 (025)*	.307 (.063)	.270 (.032)	.316 (.066)	.280 (.034)	.225
$0.5, 0.0$	.310 (.037)	.293 (.031)	**.271 (.024)**	*.251 (.001)*	.304 (.060)	.260 (.009)	.279 (.033)	.264 (.014)	.25
$0.5, 0.1$	.315 (.039)	.302 (.037)	**.271 (.018)**	*.259 (.003)*	.303 (.061)	.275 (.030)	.298 (.054)	.274 (.016)	.25

TABLE II

RESULTS (MEAN ERROR RATE AND STANDARD DEVIATION OVER 20 RANDOM EXPERIMENTS) FOR ALL TASKS. WINNERS ARE REPORTED INDEPENDENTLY FOR HOLDOUT AND EXTENDED VALIDATION EXPERIMENTS.

experiment	ν-SVM holdout	ν-SVM extended	GO-SVM holdout	GO-SVM extended	LO-SVM holdout	LO-SVM extended	SVM+ holdout	SVM+ extended
T.ser 20	.274 (.126)	.229 (.079)	**.190 (.116)**	*.149 (.101)*	.261 (.120)	.251 (.118)	.260 (.123)	.212 (.123)
T.ser 50	.106 (.035)	.092 (.029)	**.052 (.025)**	*.032 (.009)*	.089 (.040)	.075 (.030)	.058 (.025)	.042 (.017)
Surv 20	.400 (.060)	.368 (.052)	.376 (.052)	*.338 (.038)*	.380 (.053)	.364 (.050)	**.374 (.051)**	.374 (.051)
Surv 40	.327 (.030)	.315 (.026)	**.290 (.035)**	*.272 (.029)*	.313 (.029)	.302 (.029)	.345 (.034)	.345 (.034)
Surv 100	.255 (.028)	.249 (.028)	**.226 (.019)**	*.215 (.014)*	.243 (.021)	.226 (.019)	.286 (.028)	.284 (.026)
Digits 60	.114 (.034)	.110 (.029)	**.111 (.032)**	*.099 (.024)*	.114 (.036)	.107 (.027)	.113 (.032)	.113 (.032)
Digits 80	.091(.013)	.089 (.014)	**.086 (.014)**	*.077 (.010)*	.090 (.016)	.085 (.013)	.091 (.012)	.091 (.012)
Ames 100	.115 (.014)	.110 (.010)	**.096 (.010)**	*.092 (.006)*	.106 (.013)	.105 (.011)	.104 (.011)	.097 (.009)

information as a one-dimensional correcting space.[10] The author is aware that the comparison has flaws, as the specificity of assumptions proposed here gives the CPR formulations an advantage. Nevertheless, the evaluation is informative as to whether the stronger CPR assumptions yield some benefit over the more general SVM+ assumptions.

The first evaluation is up/down prediction of the MacKey-Glass synthetic timeseries [22]. It was used in the LUPI setting (SVM+) in [2], where the authors used a 4-dimensional embedding $(x_{t-3}, x_{t-2}, x_{t-1}, x_t)$ in order to predict $x_{t+5} > x_t$. In that experiment, privileged information was a 4-dimensional embedding around the target: $(x_{t+3}, x_{t+4}, x_{t+6}, x_{t+7})$. The authors compared SVM+ to SVM. Their results for both SVM and SVM+ could not be replicated.[11] We interpret $x_{t+5} - x_t$ as an increasing transformation of η. We use an RBF kernel for all experiments with this dataset. This is indicated in the results chart by T.ser.

The second evaluation is predicting binary survival at a fixed time from onset. Synthetic datasets are created using the same procedure as Shiao and Cherkassky [23], with noise level .1 and no censoring. While censored data (such as survival time known only to be greater than some value) are an inherent aspect of survival studies, it is avoided because ordinal models can be modified to accommodate the partial information that censored examples contain; this extension is an experiment for another day. This experiment uses the difference in the fixed prediction horizon and the event time for the conditional probability regularizer and considers only linear models.

The third evaluation is handwritten digit recognition, which was used by Vapnik and Vahist [2] for SVM+ and slightly adapted by Lapin *et al.* for their proposed LUPI method [4]. The task is to classify down-sampled (10×10) MNIST images based on pixel values. Lapin added human-annotated confidence scores to training examples (available for download). We repeat the experiment using their data preparation and using their annotators' confidence scores as the conditional probability regularizer. These experiments use an RBF kernel. They averaged the scores of 16 annotators, whereas we selected the grades given by annotator 6, whose labels were closest to a uniform distribution of the possible grades.

The last evaluation is based on predicting whether a house in Ames, IA, has an assessed value above or below the median price [24]. The dataset is comparable to the well-known Boston Housing UCI dataset, but with considerably greater detail about subject houses. The dataset gives sale prices and 318 attributes after converting ordinal variables to categorical indicators for 2924 houses and following the dataset author's recommendations of removing a few outliers. Empirical conditional probability is the actual assessed value.

[10]SVM+ parameters were selected as follows: The 'decision space' kernel parameter was fixed as in the other CPR formulations described previously. The RBF kernel was used for the correcting space, with kernel widths corresponding to the $[.1, .5, .9]$ quantiles of the pairwise distances—in this case, single-dimensional data. The cost parameters were chosen by grid search from $2^{-4}, \ldots, 2^{8}$, and the correcting complexity parameter (γ) was chosen from $2^{-4}, \ldots, 2^{4}$.

[11]The parameters used to generate the timeseries were an integration step size of .1, with points created every 10 steps, a, delay constant $\tau = 17$, and initial value .9.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

C. Discussion & Conclusions

In virtually every experiment, the CPR formulations have lower loss then the baseline SVM and conditional probability regression formulations. A minimal conclusion is that the CPR information can allow one to find information about the decision boundary that labels do not convey. One should remember that the uniform performance improvement is by design, as all the CPR formulations can emulate the baseline approach under the correct parameter settings. This design choice removes the possibility of doing considerably worse than the baseline by giving up some possible performance increase. With more parameters to be found the gap between an optimal model and the selected model increases. This is seen in the results tables in the difference between 'holdout' and 'extended' model selection.

Table I shows the synthetic data experiment results. Rows correspond to various dataset generation parameters θ and σ. Columns correspond to the learning formulations evaluated and to the model selection procedure. Generally, there is a trend toward greater absolute reduction in loss compared to the baselines as the problems become more difficult (as θ increases) for a fixed σ (particularly noticeable in the extended validation results). This corresponds to the observation from theory that CPR methods are unlikely to provide any benefit to classification problems that are separable or have low loss.

For $\theta = .3$ tasks (that have a comparatively low Bayes rate), the LO-SVM formulation appears to outperform GO-SVM, but otherwise the GO-SVM formulation appears stronger. Moreover, the addition of noise ($\sigma = .1$) to the conditional probability information diminishes the advantage for LO-SVM. The conditional probability noise experiments (increasing σ for $\theta = .45$) indicate that the GO-SVM formulation attains consistently lower loss than baselines under optimal ('extended') model selection at all levels of conditional probability noise, but the gap between optimal and practical model selection reduces the practical advantage. It is noteworthy that the GO-SVM extended formulations nearly achieve the Bayes optimal rates with low conditional probability noise.

Results of common dataset experiments are given at Table II. Sizes for training are given in parenthesis with the experiment name. The author points out that many LUPI research papers require validation sets that are comparable to the 'extended' experiment setting. Each row of the table gives the size of the training and test sets. The columns correspond to different methods and model selection procedures.

The GO-SVM formulation is far-and-away the best at the extended validation task. While optimal model selection is not practically attainable, the experiment is intended to illustrate how the formulations perform if model selection were assumed away. The gap between the standard holdout and extended holdout datasets is larger for the CPR formulations than SVM. It is possible that the CPR formulations, in addition to providing some hypothesis spaces with low variance in excess risk, can also have high variance. The author suspects that traditional hold-out model selection strategies are more difficult with CPR because the nested-hypothesis space assumptions of structural risk minimization [1] no longer hold. In the CPR framework, there is no total ordering of hypothesis complexity. Some hypothesis spaces (defined by parameters) have good convergence, while others do not. The task is to differentiate them. In some ways, the problem is the result of the automated nature of the experiments herein. The author believes that a human visualizing the results over the entire validation grid could make a better choice.

These results can be compared with the results of comparable experiments in other papers. The MacKey-Glass experiment appeared in the original SVM+ paper [2]. Results could not be replicated in that similar levels of performance seemed to be attainable with a smaller sample than was required there for all formulations. Both CPR and SVM+ show lower loss relative to baseline than reported there. The Digits experiment is intended to replicate one in [4]. This paper specifically replicated the experiment in which conditional probability weights were created by human annotators. This task is well-suited to the order invariance that CPR formulations enjoy, as humans have a fundamentally ordinal notion of confidence. For a sample of size 80, GO-SVM achieved a score of .077 based on extended model selection while their best method achieved .073 with a comparable large validation set. Their best method, however, did not use human confidence; rather, the human confidence information improved over the baseline SVM insignificantly.

The author observes that LO-SVM formulations were frequently primal unbounded/dual infeasible for certain settings of C. Being able to determine valid ranges for the parameters may allow a better exploration of the parameter search and improve results.

In conclusion, the results show that the methods show practical promise. However, the large parameter grid would impede any practically-sized task, especially with cross-validation. The author believes that these problems can be overcome, but considerably more study is needed.

REFERENCES

[1] V. Vapnik, *Statistical learning theory*, ser. Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998. [Online]. Available: http://books.google.com/books?id=GowoAQAAMAAJ

[2] V. Vapnik and A. Vashist, "2009 special issue: A new learning paradigm: Learning using privileged information," *Neural Netw.*, vol. 22, pp. 544–557, July 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1594405.1594439

[3] D. Pechyony and V. Vapnik, "On the theory of learning with privileged information," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010.

[4] M. Lapin, M. Hein, and B. Schiele, "Learning using privileged information: SVM+ and weighted SVM,"

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Neural Networks, vol. 53, pp. 95–108, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2014.02.002

[5] D. Lopez-Paz, B. Schölkopf, L. Bottou, and V. Vapnik, "Unifying distillation and privileged information," in *International Conference on Learning Representations (ICLR)*, Nov. 2016.

[6] J. Chen, X. Liu, and S. Lyu, "Boosting with side information," in *Computer Vision - ACCV 2012 - 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I*, 2012, pp. 563–577. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37331-2_43

[7] Z. Wang, T. Gao, and Q. Ji, "Learning with hidden information using a max-margin latent variable model," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, Aug 2014, pp. 1389–1394.

[8] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider, "Learning using privileged information in prototype based models," in *Artificial Neural Networks and Machine Learning ICANN 2012*, ser. Lecture Notes in Computer Science, A. Villa, W. Duch, P. rdi, F. Masulli, and G. Palm, Eds. Springer Berlin Heidelberg, 2012, vol. 7553, pp. 322–329. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33266-1_40

[9] Z. Wang and Q. Ji, "Classifier learning with hidden information," *Computer Vision and Pattern Recognition, IEEE Conference on*, 2015.

[10] V. Vapnik and R. Izmailov, "Rethinking statistical learning theory: learning using statistical invariants," *Mach. Learn.*, vol. 108, no. 3, pp. 381–423, 2019. [Online]. Available: https://doi.org/10.1007/s10994-018-5742-0

[11] Z. Xiao, Z. Luo, B. Zhong, and X. Dang, "Robust and efficient boosting method using the conditional risk," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, pp. 1–15, 06 2017.

[12] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *Ann. Statist.*, vol. 32, no. 1, pp. 135–166, 02 2004. [Online]. Available: http://dx.doi.org/10.1214/aos/1079120131

[13] P. L. Bartlett, M. I. Jordan, and J. D. Mcauliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006. [Online]. Available: http://www.jstor.org/stable/30047445

[14] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of u-statistics," *The Annals of Statistics*, pp. 844–874, 2008.

[15] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *The annals of mathematical statistics*, pp. 293–325, 1948.

[16] T. Vacek, "Ordering as privileged information," *CoRR*, vol. abs/1606.09577, 2016. [Online]. Available: http://arxiv.org/abs/1606.09577

[17] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *Ann. Statist.*, vol. 33, no. 4, pp. 1497–1537, 08 2005. [Online]. Available: http://dx.doi.org/10.1214/009053605000000282

[18] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification : a survey of some recent advances," *ESAIM: Probability and Statistics*, vol. 9, pp. 323–375, 2005. [Online]. Available: http://eudml.org/doc/244764

[19] B. Schölkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," in *Computational Learning Theory*, ser. Lecture Notes in Computer Science, D. Helmbold and B. Williamson, Eds. Springer Berlin Heidelberg, 2001, vol. 2111, pp. 416–426. [Online]. Available: http://dx.doi.org/10.1007/3-540-44581-1_27

[20] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, 2002, pp. 937–944.

[21] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, May 2000. [Online]. Available: http://dx.doi.org/10.1162/089976600300015565

[22] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines," in *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, Sep 1997, pp. 511–520.

[23] H. Shiao and V. Cherkassky, "Learning using privileged information (LUPI) for modeling survival data," in *2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014*, 2014, pp. 1042–1049. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2014.6889517

[24] D. DeCock, "Ames, Iowa: Alternative to the boston housing data as an end of semester regression project," *Journal of Statistics Education*, vol. 19, no. 3, 2011. [Online]. Available: www.amstat.org/publications/jse/v19n3/decock.pdf

978-1-7281-6926-2/20/$31.00 ©2020 IEEE