Non-Enumerative Cross Validation for the Determination of Structural Parameters in Feature-Selective SVMs

Elena Chernousova, Pavel Levdik
Moscow Institute of Physics and Technology, Moscow, Russia
elchernousova@inbox.ru, cold62@mail.ru

Alexander Tatarchuk, Vadim Mottl
Computing Center of the Russian Academy of Sciences, Moscow, Russia
aitech@yandex.ru, vmottl@yandex.ru

David Windridge
University of Surrey
Guildford, UK
d.windridge@surrey.ac.uk

Abstract*— The relational approach to dependency estimation entails the selection of a sufficiently compact 'relevance' subset of training-set objects with which any newly occurring object may be compared in order to estimate its hidden target characteristics. If several comparison modalities are available, a 'relevance' subset of these may additionally have to be chosen via an appropriate selection criterion. Typically, the level of selectivity will constitute a free parameter, and in traditional approaches, multiple training repetitions would be required to determine this value via cross-validation. To avoid this, we seek to algorithmically emulate the cross-validation process using conservative assumptions as to the nature of the unknown probability distribution that produced the training set. We term this approach 'non-enumerative cross-validation', and demonstrate that the classical Akaike Information Criterion is a specific case of it under naïve assumptions. The application of this non-enumerative cross-validation strategy is demonstrated on the standard multikernel data set, "chicken-pieces", treated from the perspective of relational discriminant analysis.

***Keywords*— relational dependence estimation; relevance vector machine; support vector machine; feature selection; selectivity adjustment; Akaike information criterion; non-enumerative cross-validation; non-enumerative model verification**

I. INTRODUCTION

Given a finite training set of real-world objects $\omega\in\Omega$ represented by real-valued feature vectors

$$\boldsymbol{x}(\omega)=\left(x_i(\omega),\, i\in\mathbb{I}\right)\in\mathbb{R}^n,\ n=|\mathbb{I}|,\ \mathbb{I}=\{1,...,n\}, \qquad (1)$$

and labeled by some normally hidden numerical characteristic

$$(\boldsymbol{X},\boldsymbol{y})=\left\{\left(\boldsymbol{x}(\omega_j)=\boldsymbol{x}_j, y(\omega_j)=y_j\right), y_j\in\mathbb{R}\ or\ y_j=\pm 1,\, j\in\mathbb{J}\right\}, \mathbb{J}=\{1,...,N\}, \qquad (2)$$

the linear methods of dependence estimation yield a linear decision rule in the feature space $\left(\boldsymbol{a}=(a_i, i=1,...,n)\in\mathbb{R}^n,\ b\in\mathbb{R}\right)$, which is applicable to any new object $(\boldsymbol{x},y)\notin\left\{(\boldsymbol{x}_j,y_j),\, j\in\mathbb{J}\right\}$:

$$d(\boldsymbol{x})=\boldsymbol{a}^T\boldsymbol{x}+b\begin{cases}=\hat{y}(\boldsymbol{x})-\textit{regression estimation},\\ \gtrless 0,\ \hat{y}(\boldsymbol{x})=\pm 1-\textit{pattern recognition}.\end{cases} \qquad (3)$$

We consider here primarily the problem of pattern recognition, namely, one of several versions of the commonly adopted Support Vector Machine (SVM) [1,2]:

$$\begin{cases}\sum_{i\in\hat{\mathbb{I}}} a_i^2 + C\sum_{j\in\mathbb{J}}\xi_j^2\to min\,(a_i, i\in\hat{\mathbb{I}}, b, \xi_j,\, j\in\mathbb{J}),\\ y_j\left(\sum_{i\in\hat{\mathbb{I}}} a_i x_{ij}+b\right)\geq 1-\xi_j,\ \xi_j\geq 0,\ j\in\mathbb{J}=\{1,...,N\}.\end{cases} \qquad (4)$$

Here $\hat{\mathbb{I}}\subseteq\mathbb{I}=\{1,...,n\}$ is a subset of "active" features, $\hat{n}=|\hat{\mathbb{I}}|\leq n$. If $\hat{\mathbb{I}}$ is fixed, the major advantage of SVM is that it selects a relatively small subset of training-set feature vectors $\boldsymbol{x}_j\in\mathbb{R}^{\hat{n}}$, $j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\subset\mathbb{J}$, called support vectors, which completely determine the estimated direction vector of the discriminant hyperplane $\hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}\in\mathbb{R}^{\hat{n}}$. It is easy to prove [1,2] that the estimated direction vector is a linear combination of only *support vectors*

$$\hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C}=\sum\nolimits_{j\in J} y_j\hat{\lambda}_{\hat{\mathbb{I}},C,j}\boldsymbol{x}_j\in\mathbb{R}^{\hat{n}}. \qquad (5)$$

Here the variables $(\hat{\lambda}_{\hat{\mathbb{I}},C,j},\, j\in\mathbb{J})$ are the estimated nonnegative Lagrange multipliers at the inequality constraints $y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)\geq 1-\xi_j$ in (4). If $\lambda_j=0$, the respective constraint is inactive $y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)>1-\xi_j$ and $\xi_j=0$; contrarily, when $\lambda_j>0$, this implies that the constraint is active $y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)=1-\xi_j$, and the respective training-set object is said to be a *support object*. As a result, the final decision rule is typically much simpler than the full expression (3)

$$d(\boldsymbol{x})=\sum\nolimits_{j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} y_j\hat{\lambda}_{\hat{\mathbb{I}},C,j}\boldsymbol{x}_j^T\boldsymbol{x}+\hat{b}_{\hat{\mathbb{I}},C}\gtrless 0,\ \hat{\mathbb{J}}_{\hat{\mathbb{I}},C}=\{j:\hat{\lambda}_{\hat{\mathbb{I}},C,j}>0\}\subseteq\mathbb{J}, \qquad (6)$$

and is completely defined by the subset of support vectors $\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}$ along with their class-memberships $\left\{(\boldsymbol{x}_j,y_j),\, j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\right\}$ as well as the positive Lagrange multipliers associated with them $\{\hat{\lambda}_{\hat{\mathbb{I}},C,j}>0,\, j\in\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\}$.

The choice of the active feature subset $\hat{\mathbb{I}}\subseteq\mathbb{I}$ along with the value of the structural parameter $C>0$ in (4) completely determine the number of support vectors $\hat{N}_{\hat{\mathbb{I}},C}=|\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|$, and is thus a measure of the complexity of the decision rule (6). The pair of structural parameters $(\hat{\mathbb{I}},C)$ has, hence, a critical bearing on the generalization performance of the resulting SVM.

If no separate test set is available, the only way to adjust $(\hat{\mathbb{I}},C)$ is via cross validation within the training set (2). However, traditional *explicit* cross-validation requires multiple training repetitions to adjust the structural parameters $(\hat{\mathbb{I}},C)$.

In this paper, we provide a mathematical justification of the suggestion made in [3] that the Akaike Information Criterion (AIC) [4], originally developed as applied to regression estimation, can be viewed as hypothetical cross-validation. It has the advantage that an analytical expression can be obtained for comparing models. We proceed from a more general view of machine learning, and mathematically emulate the cross-validation process by exploiting certain conservative assumptions regarding the probability distribution giving rise to the training set. We call this principle "*hypothetical non-enumerative cross-validation*", and show that the classical AIC constitutes a particular case under certain assumptions. Its application to SVMs explicitly exploits the dependency of training results on the existence of support feature vectors.

We apply the principle of non-enumerative hypothetical cross-validation not only to the classical SVM, with a fixed set of object features $\hat{\mathbb{I}}\subseteq\mathbb{I}$ (1), but also to the Relevance Vector

Machine (RVM) [5,6,8] based on the assumption that the real-world objects $\omega \in \Omega$ are perceptible only by an arbitrary measure $S(\omega', \omega'')$ of their pair-wise similarity or dissimilarity. The idea is to treat the values of this function between an arbitrary object $\omega \in \Omega$ and those of the training set $\{\omega_j, j=1,...,N\}$ as the vector of secondary features

$$\boldsymbol{x}(\omega) = \left(x_i(\omega) = S(\omega_i, \omega), i \in \mathbb{I} = \mathbb{J}\right), \ \mathbb{I} = \{1,...,n\} = \{1,...,N\}, \quad (7)$$

and apply then the standard SVM in $\mathbb{R}^n = \mathbb{R}^N$. We consider a feature-selective generalization of SVM

$$\begin{cases} \sum_{i\in\mathbb{I}} a_i^2 + \mu \sum_{i\in\mathbb{I}} |a_i| + C \sum_{j\in J} \xi_j^2 \to \min(a_i, i \in \mathbb{I} = \mathbb{J}, b, \delta_j, j \in \mathbb{J}), \\ y_j \left(\sum_{i\in\mathbb{I}} a_i x_{ij} + b \right) \geq 1 - \xi_j, \xi_j \geq 0, \ j \in \mathbb{J} = \{1,...,N\}, \end{cases} \quad (8)$$

which differs from that outlined in [7] under the name of *Doubly Regularized SVM* or *Elastic Net SVM* only by the squared penalty ξ_j^2, instead of ξ_j, for violation of the generic requirement of SVM to provide a positive margin between two classes of training-set objects $y_j \left(\sum_{i\in\mathbb{I}} a_i x_{ij} + b \right) \geq 1$.

The presence of the L_1 regularization term with weighting parameter $\mu \geq 0$, as distinct from (4), yields the intrinsic property of the doubly regularized SVM to assign strictly zero values to redundant elements of the direction vector a_i, thereby automatically finding the subset of informative secondary features $\hat{\mathbb{I}}_{C\mu} = \{\hat{a}_{C\mu,i} \neq 0\} \subset \mathbb{I} = \mathbb{J}$, namely, of the training-set objects.

These objects are said to be *relevance objects*, or *relevance vectors* if the comparison function $S(\omega', \omega'')$ possesses the properties of a kernel that embeds the objects into a linear space [6]. If $\mu = 0$, the method equates to the classical SVM retaining all the features $\hat{\mathbb{I}}_{C\mu} = \mathbb{I} = \mathbb{J}$. Alternatively, if $\mu \to \infty$, the criterion becomes excessively selective $\hat{\mathbb{I}}_{C\mu} \to \varnothing$. Thus, as the structural parameter μ grows, the training ranges from the full conservation of secondary features to extreme feature selectivity.

This fact is essentially exploited by our hypothetical non-enumerative cross-validation, allowing us to avoid very computationally-expensive explicit cross validation when adjusting the structural parameters (C, μ), primarily, the selectivity parameter μ, and providing, therethrough, the best generalization performance of the doubly regularized SVM.

II. The Principle of Hypothetical Cross Validation

Our principle of non-enumerative hypothetical cross validation is based on two heuristics to be formulated in Subsections II.B, II.C and II.D. To clarify terminology, we shall first briefly outline, in Subsection II.A, the general problem of dependence estimation from empirical data.

A. The general problem of dependence estimation

Let Ω be a set of real-world objects $\omega \in \Omega$ each of which is associated with two measurable characteristics in arbitrary domains $\boldsymbol{x}(\omega) \in \mathbb{X}$ and $y(\omega) \in \mathbb{Y}$, the former of which is given and the latter hidden during the test phase. Objects $\omega \in \Omega$ are assumed to be repeatedly and independently drawn from a preexisting distribution, i.e., as a pair $(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}$. The distribution density

$$\int_{\mathbb{Y}} \int_{\mathbb{X}} f^*(\boldsymbol{x}, y) d\boldsymbol{x} dy = 1, \ f^*(x, y) = \mathrm{g}^*(\boldsymbol{x}) \varphi^*(y | \boldsymbol{x}) \quad (9)$$

is unknown to the observer who wishes to solve the problem $\hat{y}(x): \mathbb{X} \to \mathbb{Y}$, i.e., estimate the hidden dependence.

In particular, in the pattern-recognition problem $\mathbb{Y} = \{-1, 1\}$, the adoption of a conditional distribution density $\varphi^*(y|x)$ in (9) is consistent with the binary essence of the class-membership index $y = \pm 1$ because it is possible, in the case of two classes, to treat this as a real-valued random variable concentrated in two points and described by a singular density

$$\varphi^*(y | \boldsymbol{x}) = (1 - p^*) \delta\big(y - (-1)\big) + p^* \delta(y - 1), \quad (10)$$

where $\delta(z)$ is Dirac delta function, $p^* = P^*(y = 1 | \boldsymbol{x})$.

Suppose the observer has obtained a finite set of independent random drawings, i.e., a training set $(\boldsymbol{X}, \boldsymbol{y})$ as in (2). Further suppose that the observer proposes to employ a parametric class of decision rules $\hat{y}(\boldsymbol{x}, \boldsymbol{a})$, $\boldsymbol{a} \in \mathbb{R}^n$, and a loss function

$$q(y, \boldsymbol{x}, \boldsymbol{a}), \text{ for instance, } q(y, \boldsymbol{x}, \boldsymbol{a}) = Loss\big(y, \hat{y}(\boldsymbol{x}, \boldsymbol{a})\big). \quad (11)$$

The optimal way to select parameter $\boldsymbol{a}$ would be via minimization of the average risk $\int_{\mathbb{X}} \int_{\mathbb{Y}} q(y, \boldsymbol{x}, \boldsymbol{a}) f^*(\boldsymbol{x}, y) dy d\boldsymbol{x} \to \min(\boldsymbol{a})$, however this is impossible because the distribution is unknown. The commonly adopted compromise is to minimize the empirical risk computed from the available training set

$$Q(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{a}) = \sum_{j=1}^{N} q(y_j, \boldsymbol{x}_j, \boldsymbol{a}) \to \min(\boldsymbol{a}). \quad (12)$$

Let, further, the observer have a quite vague *a priori* suggestion on the value of the main parameter $\boldsymbol{a} \in \mathbb{R}^n$, which is expressed in the form of a parametric family of functions to be minimized $V(\boldsymbol{a}, C) \to \min(\boldsymbol{a})$. Here C is an additional scalar or vector parameter, constituting the so-called *structural parameter*, meant to control the undesirability of the deflection of $\boldsymbol{a}$ from a subset $\mathbb{A} \subset \mathbb{R}^n$ associated with "especially simple" decision rules.

It is common practice to accept a trade-off training criterion

$$\hat{\boldsymbol{a}}_C(\boldsymbol{X}, \boldsymbol{y}) = \arg\min_{\boldsymbol{a}} \{V(\boldsymbol{a}, C) + Q(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{a})\} \quad (13)$$

to give a regularized version. It is clear that the result of training will depend critically on the value of the structural parameter C, and its choice is, perhaps, the central problem of the machine learning theory.

B. The assumption of the tractability of the learning problem

It is always possible to represent the unknown joint probability density of the hidden and observable characteristics of a random real-world object $f^*(\boldsymbol{x}, y)$ as product of the marginal density of one variable and the conditional density of the other (9). Thus, the joint probability density of the training set as a whole can be represented as a product

$$F^*(\boldsymbol{X}, \boldsymbol{y}) = G^*(\boldsymbol{X}) \Phi^*(\boldsymbol{y} | \boldsymbol{X}) = \prod_{j=1}^{N} g^*(\boldsymbol{x}_j) \varphi^*(y_j | \boldsymbol{x}_j). \quad (14)$$

Of course, both densities remain unknown here, but let the observer try to slightly temper his/her despair of complete ignorance, and mentally tie the nature's conditional distribution $\varphi^*(y | \boldsymbol{x})$ to the parameter $\boldsymbol{a}$ that exists only in the ob-

server's imagination $\varphi^*(y|\boldsymbol{x})=\int_{\mathbb{R}^n}\varphi(y|\boldsymbol{x},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}$. In terms of the assumed mechanism of forming the training set, this means the equality

$$\Phi^*(\boldsymbol{y}|\boldsymbol{X})=\int_{\mathbb{R}^n}\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\text{, where}$$

$$\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})=\prod_{j=1}^{N}\varphi(y_j|\boldsymbol{x}_j,\boldsymbol{a}) \quad (15)$$

is treated as the completely known parametric family of conditional distributions, whereas $\Psi^*(\boldsymbol{a})$, on the contrary, is assumed to be absolutely unknown. In other words, the observer considers $\Phi^*(\boldsymbol{y}|\boldsymbol{X})$ in (14) as an unknown parametric mixture of known conditional distributions.

The treatment of the parametric family $\varphi(y|\boldsymbol{x},\boldsymbol{a})$ via the exponential of the loss function

$$\varphi(y|\boldsymbol{x},\boldsymbol{a})\propto\exp\bigl(-q(y,\boldsymbol{x},\boldsymbol{a})\bigr),\ \Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\propto\exp\bigl(-Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})\bigr), \quad (16)$$

where the normalization coefficient $\propto$ does not depend on $\boldsymbol{a}$, is equivalent to an assumption of the *non-viciousness of nature*, i.e. it is implicitly assumed, for each $\boldsymbol{a}\in\mathbb{R}^n$, that pairs $(\boldsymbol{x},y)$ corresponding to low values of the accepted loss function $q(y,\boldsymbol{x},\boldsymbol{a})$ are produced more frequently than for high values.

C. Mental experiment

Suppose, firstly, that a value of the hypothetical parameter $\boldsymbol{a}\in\mathbb{R}^n$ has been randomly drawn by the nature in accordance with the unknown density $\Psi^*(\boldsymbol{a})$, as well as all the observable characteristics of the training-set objects $\boldsymbol{X}=(\boldsymbol{x}_1,...,\boldsymbol{x}_N)$ in accordance with equally unknown density $G^*(\boldsymbol{X})$ (14).

Suppose, further, that we imagine that we have randomly and independently drawn two versions of the object characteristics $\boldsymbol{y}=(y_1,...,y_N)$ and $\tilde{\boldsymbol{y}}=(\tilde{y}_1,...,\tilde{y}_N)$. There are thus now two different hypothetical sets $(\boldsymbol{X},\boldsymbol{y})$ and $(\boldsymbol{X},\tilde{\boldsymbol{y}})$ with the same values of $\boldsymbol{X}=(\boldsymbol{x}_1,...,\boldsymbol{x}_N)$. We can imagine that these are used as the training set $(\boldsymbol{X},\tilde{\boldsymbol{y}})$, which yields some estimate of the goal parameter $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})$ (13), and the test set $(\boldsymbol{X},\boldsymbol{y})$, which is used for computing the loss $Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\bigr)$. The essence of the hypothetical cross validation is minimization of the mathematical expectation of the loss:

$$\int\!\!\int\!\!\int Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\bigr)\Bigl(\int\Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})d\boldsymbol{a}\Bigr)\times \\ \times G^*(\boldsymbol{X})d\,\tilde{\boldsymbol{y}}\,d\,\boldsymbol{y}\,d\,\boldsymbol{X}\to\min(C). \quad (17)$$

In reality, we have one training set $(\boldsymbol{X},\boldsymbol{y})$, and can only compute the loss on the same set already used for training. In this case we need to determine: 1) How large will be the defect of the criterion subject to minimization in accordance to (17)? 2) What should be the penalty for using the estimate $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})$ computed from the same set instead of an independent estimate $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})$?

Theorem 1. The equivalent form of criterion (17) is

$$C^*=\arg\min_C\Bigl\{\int_{\mathbb{X}\times\mathbb{Y}}Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{y},\boldsymbol{X})\bigr)F^*(\boldsymbol{X},\boldsymbol{y})d\boldsymbol{X}d\boldsymbol{y}+ \\ \int_{\mathbb{X}\times\mathbb{R}^m}\Delta(C,\boldsymbol{X},\boldsymbol{a})\Psi^*(\boldsymbol{a})G^*(\boldsymbol{X})d\boldsymbol{X}d\boldsymbol{c}\Bigr\}, \quad (18)$$

$$\text{where }\Delta(C,\boldsymbol{X},\boldsymbol{a})=\int_{\mathbb{Y}\times\mathbb{Y}}Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_\beta(\boldsymbol{X},\tilde{\boldsymbol{y}})\bigr)-Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\bigr)\times \\ \times\Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}d\boldsymbol{y}. \quad (19)$$

For many typical loss functions $Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})$ and regularization functions $V(\boldsymbol{a},C)$ (13) applicable to a wide class of practical problems, the penalty (19) does not depend on the parameter $\boldsymbol{a}$:

$$\int_{\mathbb{Y}\times\mathbb{Y}}Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\bigr)-Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\bigr)\times \\ \times\Phi(\tilde{\boldsymbol{y}}|\boldsymbol{X},\boldsymbol{a})\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})d\tilde{\boldsymbol{y}}d\boldsymbol{y}=\Delta(C,\boldsymbol{X}). \quad (20)$$

Theorem 2. In the case of parameter-independent penalty (19), the idea of hypothetical cross validation (17) lends itself to the simple representation:

$$C^*=\arg\min_C\Bigl\{\int\!\!\int\Bigl[Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\bigr)+\Delta\bigl(C,\boldsymbol{X}\bigr)\Bigr]F^*(\boldsymbol{X},\boldsymbol{y})d\boldsymbol{y}d\boldsymbol{X}\Bigr\}. \quad (21)$$

D. The criterion of hypothetical cross validation

However, the criterion (21) is still unfit for practical use, because the joint probability distribution is unknown to the observer. The second heuristic idea is to substitute the mathematical expectation (21) for its unbiased estimate:

$$\hat{C}(\boldsymbol{y},\boldsymbol{X})=\arg\min_C\underbrace{\Bigl\{\underbrace{Q\bigl(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\bigr)}_{\textit{empirical risk}}+\underbrace{\Delta(C,\boldsymbol{X})}_{\textit{penalty}}\Bigr\}}_{\textit{structural risk}}. \quad (22)$$

This is just the criterion of hypothetical non-enumerative cross validation we consider in this paper.

Its structure is analogous to the Vapnik-Chervonenkis criterion of structural risk minimization [1], but differs from it in the interpretation of the penalty $\Delta(C,\boldsymbol{X})$ (20). In Vapnik-Chervonenkis theory, the penalty characterizes the upper bound of the unknown average risk, which is derived from general inequalities of the probability theory and parameterized by the VC-dimension. It should be remembered that the notion of VC-dimension was formulated only for the simplest binary loss function in pattern recognition, and is inapplicable, for instance, to SVM. In contrast to this, the penalty $\Delta(C,\boldsymbol{X})$ (20) is applicable to a more wide class of loss functions, but is underlain by a potentially more restrictive heuristic assumption regarding the data. As we shall see below in Section IV, it is compatible with the SVM framework.

In accordance with the first heuristic assumption (16), we have $Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})=-\ln\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})+const$. The hypothetical cross validation $\int\bigl[\ln\Phi\bigl(\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\bigr)\bigr]\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})d\boldsymbol{y}\to\max$ thus amounts to *maximizing the Kullback information* on the unknown distribution $\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})$ contained in the estimate from another sample $\Phi\bigl(\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})\bigr)$. Therefore, it is appropriate to regard our criteria of hypothetical cross validation as implicitly information-theoretic and consider it as a generalization of the classical idea of Hirotugu Akaike set out in [4].

III. Simplest Instantiation of the Method: Linear Regression and the Akaike Information Criterion

Linear regression. Let the unobservable variable take values along the real axis $y\in\mathbb{Y}=\mathbb{R}$, and the observable one be a real vector, i.e., $\boldsymbol{x}\in\mathbb{X}=\mathbb{R}^n$. We shall assume the loss function (11)-(12) to be linear and quadratic

$$q(y,\boldsymbol{x},\boldsymbol{a})=(y-\boldsymbol{x}^T\boldsymbol{a})^2,\ \boldsymbol{X}=(\boldsymbol{x}_1\cdots\boldsymbol{x}_N)\ (n\times N), \\ Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a})=\bigl\|\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\bigr\|^2=(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a})^T(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}), \quad (23)$$

thus, the assumed conditional distribution (16) to be normal $\varphi(y|\boldsymbol{x},\boldsymbol{a})=\mathcal{N}(y|\boldsymbol{x}^T\boldsymbol{a},\sigma^2)$ with fixed variance $\sigma^2=1/2$. This is hence the problem of linear regression estimation.

In the case of the simplest quadratic regularization function $V(\boldsymbol{a},C)=\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a}$, where symmetric positive semidefinite matrix $\boldsymbol{B}_C$ $(n\times n)$ depends on the structural parameter C, the trade-off training criterion (13) will yield the estimated vector of regression coefficients

$$\hat{\boldsymbol{a}}_C(\mathbf{X},\mathbf{y})=\arg\min_{\boldsymbol{a}}\left\{\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a}+\left\|\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{a}\right\|^2\right\}=\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\boldsymbol{X}\boldsymbol{y}. \quad (24)$$

Theorem 3. For the linear-quadratic loss function (23) and, hence, the normal conditional density of the hidden variable $\Phi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{a})$ (16), the penalty (19) for using the estimate $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})$ (24) computed from the same set instead of an independent estimate $\hat{\boldsymbol{a}}_C(\boldsymbol{X},\tilde{\boldsymbol{y}})$ does not depend on the unknown parameter $\boldsymbol{a}$ (20):

$$\Delta(C,\boldsymbol{X})=Tr\left[\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\right]. \quad (25)$$

Thus, the criterion of hypothetical cross validation (22) for the linear regression model has the form

$$\hat{C}(\boldsymbol{y},\boldsymbol{X})=\arg\min_{C}\left\{\left\|\boldsymbol{y}-\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\right\|^2+Tr\left[\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\right]\right\}. \quad (26)$$

The Akaike information criterion. Let us additionally assume that the elements of the vector of regression coefficients are *a priori* ordered $\boldsymbol{a}=(a_1\cdots a_n)$, and that the integer structural parameter $0\le C\le n$ corresponds to the number of non-zero regression coefficients:

$$\boldsymbol{a}=(\boldsymbol{a}_C,\boldsymbol{a}_{n-C})=(a_1,...,a_C,a_{C+1}=0,...,a_n=0).$$

This assumption can be expressed as the simplest quadratic regularization function

$$V(\boldsymbol{a},C)=\boldsymbol{a}^T\boldsymbol{B}_C\boldsymbol{a},\ \ \boldsymbol{B}_C=Diag\left(\underbrace{\frac{1}{\rho}\cdots\frac{1}{\rho}}_{C}\underbrace{\rho\cdots\rho}_{n-C}\right),\ \rho\to\infty. \quad (27)$$

Theorem 4. Under the assumption (27)

$$\lim_{\rho\to\infty}\Delta(C,\boldsymbol{X})=\lim_{\rho\to\infty}Tr\left[\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{B}_C)^{-1}\right]=C,$$

Thus, the criterion of hypothetical cross validation (26) reduces to an especially simple form:

$$\hat{C}(\boldsymbol{y},\boldsymbol{X})=\arg\min_{C}\left\{\left\|\boldsymbol{y}-\boldsymbol{X}^T\hat{\boldsymbol{a}}_C(\boldsymbol{X},\boldsymbol{y})\right\|^2+C\right\}. \quad (28)$$

This is just the idea of the Akaike information criterion [4].

Comparison of (28) and (22) allows us to interpret the penalty term $\Delta(C,\mathbf{X})$ in the criterion of hypothetical cross validation as a *generalized real-valued dimensionality of the data model.*

IV. Second Instantiation: The Support Vector Machine

A. The subset of support vectors as a self-contained non-numeric structural parameter

Let us consider a parametric family of discriminant hyperplanes $\boldsymbol{a}^T\boldsymbol{x}+b\gtrless 0$ in an $\hat{n}$-dimensional feature space (4)

$$\boldsymbol{x}=(x_i,i\in\hat{\mathbb{I}})\in\mathbb{R}^{\hat{n}},\boldsymbol{a}=\left(a_i,i\in\hat{\mathbb{I}}\subseteq\mathbb{I}=\{1,...,n\}\right)\in\mathbb{R}^{\hat{n}},b\in\mathbb{R},|\hat{\mathbb{I}}|=\hat{n}, \quad (29)$$

Which distinguishes between two classes of objects $y=\pm 1$ represented by feature vectors $\boldsymbol{x}\in\mathbb{R}^{\hat{n}}$. The goal of classification is then to select a hyperplane such that the feature vectors of objects of different classes would fall primarily in different half-spaces. Let the loss function $q(y,\boldsymbol{x},\boldsymbol{a},b)$ (11) applicable to any object $(\boldsymbol{x},y)$ be chosen in the form

$$q(y,\boldsymbol{x},\boldsymbol{a},b)=\begin{cases}0, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b)>1,\\ \left(1-y(\boldsymbol{a}^T\boldsymbol{x}+b)\right)^2, & \text{if } y(\boldsymbol{a}^T\boldsymbol{x}+b)<1,\end{cases} \quad (30)$$

which penalizes the feature vector $\boldsymbol{x}$ if it is located on the wrong side of the hyperplane, but also penalizes proximities on the correct side of the hyperplane of less than 1 assuming a Euclidean metric. The respective empirical risk of the training set (2) will be the sum

$$Q(\boldsymbol{y},\boldsymbol{X},\boldsymbol{a},b)=\sum_{j\in\mathbb{J}}q(y_j,\boldsymbol{x}_j,\boldsymbol{a},b)=\sum_{j:y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1}\left(y_j-(\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2. \quad (31)$$

If we assume the quadratic regularization function $V(\boldsymbol{a},b)=\boldsymbol{a}^T\boldsymbol{a}$, then training criterion (13) will have the form

$$\begin{pmatrix}\hat{\boldsymbol{a}}_{\mathbb{I},C}(\boldsymbol{X},\boldsymbol{y})\\ \hat{b}_{\mathbb{I},C}(\boldsymbol{X},\boldsymbol{y})\end{pmatrix}=\arg\min_{\boldsymbol{a},b}\left\{\boldsymbol{a}^T\boldsymbol{a}+C\sum_{j:\,y_j(\boldsymbol{a}^T\boldsymbol{x}_j+b)<1}\left(y_j-(\boldsymbol{a}^T\boldsymbol{x}_j+b)\right)^2\right\}, \quad (32)$$

which equates to the standard SVM criterion (4).

It will be convenient to us to eliminate the double notation of the parameters of the discriminant hyperplane $(\boldsymbol{a},b)$ by adding an extra element 1 to the feature vector and an extra b to the direction vector, so that $\mathbf{x}=(\boldsymbol{x},1)\in\mathbb{R}^{\hat{n}+1}$, $\mathbf{a}=(\boldsymbol{a},b)\in\mathbb{R}^{\hat{n}+1}$. In this case, the SVM problem (4) will have the form

$$\begin{cases}\mathbf{a}^T\mathbf{B}_{\hat{\mathbb{I}},C}\mathbf{a}+\sum_{j\in\mathbb{J}}\xi_j^2\to min\,(\mathbf{a},\xi_j,\ j\in\mathbb{J}),\\ y_j\mathbf{a}^T\mathbf{x}_j\ge 1-\xi_j,\ \xi_j\ge 0,\ j\in\mathbb{J}=\{1,...,N\},\end{cases}\ \mathbf{B}_{\hat{\mathbb{I}},C}=\underset{((\hat{n}+1)\times(\hat{n}+1))}{\begin{pmatrix}(1/C)\boldsymbol{I}_{\hat{n}} & \mathbf{0}\\ \mathbf{0}^T & 0\end{pmatrix}}. \quad (33)$$

The solution of the respective dual problem yields the optimal discriminant hyperplane $\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})$, i.e., $(\hat{\boldsymbol{a}}_{\hat{\mathbb{I}},C},\hat{b}_{\hat{\mathbb{I}},C})$, which is applicable to feature vectors of new objects (3), and the subset of support objects of the training set $\hat{\mathbb{J}}_C\subseteq\mathbb{J}$:

$$\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}=\left\{j\in\mathbb{J}:\,y_j\mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})=1-\hat{\xi}_j,\,\hat{\xi}_j>0\right\},\ \hat{\xi}_j^2=(y_j-\mathbf{x}_j^T\hat{\mathbf{a}}_{\hat{\mathbb{I}},C})^2. \quad (34)$$

Traditionally, to choose the appropriate values of the structural parameters $(\hat{\mathbb{I}},C)$, the user has to repeat training (33) for a series of tentative values, and accept the result $\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})$ that provides the minimum cross-validation error.

In order to make the notion of hypothetical non-enumerative cross validation applicable to SVM, we associate the value of the numerical structural parameter $(\hat{\mathbb{I}},C)$ with the resulting subset of support vectors $\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}\subseteq\mathbb{J}$. Only the feature vectors and class indices of the support objects affect the result of training:

$$\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}=\begin{pmatrix}\boldsymbol{x}_{j_1}\cdots\boldsymbol{x}_{j_{\hat{N}_{\hat{\mathbb{I}},C}}}\\ 1\ \cdots\ 1\end{pmatrix}\left((\hat{n}+1)\times\hat{N}_{\hat{\mathbb{I}},C}\right),\hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}=(y_{j_1}\cdots y_{j_{\hat{N}_{\hat{\mathbb{I}},C}}})^T,\ \hat{N}_{\hat{\mathbb{I}},C}=|\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|. \quad (35)$$

In accordance with (33), (34) and (35), the subset of support vectors completely defines the direction vector of the optimal discriminant hyperplane:

$$\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y})=\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X},\boldsymbol{y},\hat{\mathbb{J}}_{\hat{\mathbb{I}},C})=\text{argmin}\left\{\mathbf{a}^T\mathbf{B}_{\hat{\mathbb{I}},C}\mathbf{a}+\sum_{j\in\hat{\mathbb{J}}_\beta}(y_j-\mathbf{x}_j^T\mathbf{a})^2\right\}=$$
$$\text{argmin}\left(\mathbf{a}^T\mathbf{B}_{\hat{\mathbb{I}},C}\mathbf{a}+\left\|\hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}-\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T\mathbf{a}\right\|^2\right)=(\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T+\mathbf{B}_{\hat{\mathbb{I}},C})^{-1}\hat{\mathbf{X}}_{\hat{\mathbb{I}},C}\hat{\boldsymbol{y}}_{\hat{\mathbb{I}},C}. \quad (36)$$

We treat this subset as a self-contained structural parameter of the decision rule. Our approach is hence to consider the subset $\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}$

as the structural parameter which is subject to cross validation in a tentative series $\left[(\hat{\mathbb{I}}_1, C_1), ..., (\hat{\mathbb{I}}_m, C_m)\right] \Rightarrow \left[\hat{\mathbb{J}}_{\hat{\mathbb{I}}_1, C_1}, ..., \hat{\mathbb{J}}_{\hat{\mathbb{I}}_m, C_m}\right]$.

B. Application of a heuristic tractability assumption

By mathematical formulation, the problem (36) seems to coincide with that of regression estimation (24). However, the fundamental distinction is that the variables y_j take in (36) only two values $y_j = \pm 1$, whereas in (24) these are real variables $y_j \in \mathbb{R}$. As a consequence, an attempt to apply the assumption (16) results in inevitable dependence of the normalization coefficient $\propto$ on the unknown value of $\boldsymbol{a}$, and the subsequent mathematical framework in sections II.B and II.C becomes inapplicable.

The heuristic way out we propose here is to maximally exploit the formal analogy between (36) and (24), and literally treat the SVM problem with the fixed subset of support objects as though it would be that of regression estimation. Such a substitution leads to usage of expression (25) as the penalty for using the estimate $\hat{\mathbf{a}}_{\hat{\mathbb{I}},C}(\boldsymbol{X}, \boldsymbol{y})$ (36) computed from the same set instead of an independent estimate $\hat{\mathbf{a}}_C(\boldsymbol{X}, \tilde{\boldsymbol{y}})$:

$$\Delta(\hat{\mathbb{I}}, C, \boldsymbol{X}) = Tr\left[\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T (\hat{\mathbf{X}}_{\hat{\mathbb{I}},C} \hat{\mathbf{X}}_{\hat{\mathbb{I}},C}^T + \mathbf{B}_{\hat{\mathbb{I}},C})^{-1}\right]. \quad (37)$$

Theorem 5. $\lim_{C \to \infty} \Delta(\hat{\mathbb{I}}, C, \boldsymbol{X}) = \min\{\hat{n}, \hat{N}_{\hat{\mathbb{I}},C}\} + 1$, where $\hat{n} = |\hat{\mathbb{I}}|$, $\hat{N}_{\hat{\mathbb{I}},C} = |\hat{\mathbb{J}}_{\hat{\mathbb{I}},C}|$ in accordance with (29) and (35).

The latter theorem shows that the penalty (37) should be understood as *effective dimension* of the hyperplane's parameter $(\boldsymbol{a} \in \mathbb{R}^{\hat{n}}, b \in \mathbb{R})$, which is often smaller than $\hat{n} + 1 = |\hat{\mathbb{I}}| + 1$.

As applied to the choice of the feature subset $\hat{\mathbb{I}} \subseteq \mathbb{I}$ with a sufficiently large value of C, our criterion of hypothetical cross validation for SVM (4) and (33), in accordance with notations (35), becomes the form:

$$\hat{\mathbb{I}}_C(\boldsymbol{y}, \boldsymbol{X}) = \arg\min_{\hat{\mathbb{I}} \subseteq \mathbb{I}} \left\{ \sum_{j \in \hat{\mathbb{J}}_{\hat{\mathbb{I}},C}} \hat{\xi}_{\hat{\mathbb{I}},C,j}^2 + \left(\min\{\hat{n}, \hat{N}_{\hat{\mathbb{I}},C}\} + 1\right) \right\}. \quad (38)$$

However, the huge number 2^n of all the feature subsets $\hat{\mathbb{I}} \subseteq \mathbb{I} = \{1, ..., n\}$ prevents direct usage of this "naive" rule.

C. The Relevance Vector (Object) Machine

The additional regularization term $\mu \sum_{i \in \mathbb{I}} |a_i|$ in (8) gives rise to characteristics that are significantly different from the standard SVM. This term serves to automatically select the most informative subset of secondary features, *relevance objects* $\hat{\mathbb{I}}_{C\mu} \subset \mathbb{I} = \mathbb{J}$, whose role has much in common with that of support objects (vectors) in the classical SVM in that only *relevance objects* are associated with the non-zero coefficients at the solution of the convex training problem. Since the doubly regularized criterion is convex, it does not matter, for the outcome, which algorithm is used to obtain the solution. In particular, the algorithm proposed in [8] efficiently determines the optimal subset of secondary features (relevance objects) $\hat{\mathbb{I}}_{C\mu}$.

Heuristic evaluation has indicated that it is reasonable to fix the parameter C at a sufficiently large value while varying only the selectivity parameter μ.

Once the subset of secondary features is found

$$\hat{\mathbb{I}}_{C\mu} \subset \mathbb{I}, \ \hat{n}_{C\mu} = |\hat{\mathbb{I}}_{C\mu}|, \quad (39)$$

it appears expedient to apply the usual SVM (4) to this subset:

$$\begin{cases} \sum_{i \in \hat{\mathbb{I}}_{C\mu}} a_i^2 + C \sum_{j \in \mathbb{J}} \xi_j^2 \to min\,(a_i, i \in \hat{\mathbb{I}}_{C\mu}, b, \xi_j, j \in \mathbb{J}), \\ y_j \left(\sum_{i \in \hat{\mathbb{I}}_{C\mu}} a_i x_{ij} + b \right) \geq 1 - \xi_j, \ \xi_j \geq 0, \ j \in \mathbb{J} = \{1, ..., N\}. \end{cases} \quad (40)$$

Application of any standard training algorithm yields the subset of support objects

$$\hat{\mathbb{J}}_{C\mu} = \{j : \hat{\xi}_{C\mu,j} > 0\}, \ \hat{N}_{C\mu} = |\hat{\mathbb{J}}_{C\mu}|, \quad (41)$$

i.e., such objects that $y_j \left(\sum_{i \in \hat{\mathbb{I}}_{C\mu}} \hat{a}_{C\mu,i} x_{ij} + \hat{b}_{C\mu} \right) = 1 - \hat{\xi}_{C\mu,j}$. As distinct of (6), the subset of support objects will, first, depend on both structural parameters (C, μ), and, second, define the decision rule that takes into account only the relevance secondary features of any new object (7)

$$d(\boldsymbol{x}(\omega) | C, \mu) = d(x_1(\omega), ..., x_N(\omega) | C, \mu) = \sum_{j \in \hat{\mathbb{J}}_{C\mu}} y_j \hat{\lambda}_{C,j} \sum_{i \in \hat{\mathbb{I}}_{C\mu}} x_{ij} x_i + \hat{b}_C \gtrless 0, \ x_i = x_i(\omega) = S(\omega_i, \omega), \ i \in \mathbb{I} = \mathbb{J}. \quad (42)$$

The training at each of the points $(\mu_1 < ... < \mu_m)$ yields the respective succession of the relevance sets $(\hat{\mathbb{I}}_{C\mu_1}, ..., \hat{\mathbb{I}}_{C\mu_m})$ of sizes $(\hat{n}_{C\mu_1}, ..., \hat{n}_{C\mu_m})$, which generally show the tendency to form diminishing subsets $(\hat{\mathbb{I}}_{C\mu_1} \supset \hat{\mathbb{I}}_{C\mu_2} \supset ... \supset \hat{\mathbb{I}}_{C\mu_m})$, however, the latter characteristic is not always strongly evident. It is in this context that we wish to determine the most appropriate selectivity setting via hypothetical cross validation.

Our treatment of the doubly regularized SVM (8), which we regard as a Relevance Vector Machine with supervised selectivity, is eligible to the same heuristic trick that we applied to the usual SVM in section IV.B. The penalty (37) for the incorrect estimate is completely applicable to the SVM formulation (40) with the fixed subset of secondary features $\hat{\mathbb{I}}_{C\mu} \subset \mathbb{I} = \mathbb{J}$ previously estimated by the doubly regularized SVM (8). The only distinction is that matrices $\mathbf{B}_{C\mu}$ and $\hat{\mathbf{X}}_{C\mu}$ are now of smaller dimension than (33) and (35) in accordance with (39) and (41):

$$\hat{\mathbf{B}}_{C\mu} = \begin{pmatrix} (1/C)\boldsymbol{I}_{\hat{n}_{C\mu}} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix} \left((\hat{n}_{C\mu} + 1) \times (\hat{n}_{C\mu} + 1) \right),$$

$$\hat{\mathbf{X}}_{C\mu} = \begin{pmatrix} \hat{\boldsymbol{x}}_{C\mu,j_1} ... \hat{\boldsymbol{x}}_{C\mu,j_{\hat{N}_{C\mu}}} \\ 1 \ \ ... \ \ 1 \end{pmatrix} \left((\hat{n}_{C\mu} + 1) \times \hat{N}_{C\mu} \right), \hat{\boldsymbol{x}}_{C\mu,j} = (x_{i_1,j} ... x_{i_{\hat{n}_{C\mu}},j})^T \in \mathbb{R}^{\hat{n}_{C\mu}}.$$

Respectively, the criterion of hypothetical cross-validation (38), as applied to the choice of μ with fixed C, will take the form

$$\hat{\mu}(\boldsymbol{y}, \boldsymbol{X}) = \arg\min_{\mu} \left\{ \sum_{j \in \hat{\mathbb{J}}_{C\mu}} \hat{\xi}_{C\mu,j}^2 + Tr\left[\hat{\mathbf{X}}_{C\mu} \hat{\mathbf{X}}_{C\mu}^T (\hat{\mathbf{X}}_{C\mu} \hat{\mathbf{X}}_{C\mu}^T + \mathbf{B}_{C\mu})^{-1} \right] \right\} \cong \arg\min_{\mu} \left\{ \sum_{j \in \hat{\mathbb{J}}_{C\mu}} \hat{\xi}_{C\mu,j}^2 + \left(\min\{\hat{n}_{C\mu}, \hat{N}_{C\mu}\} + 1 \right) \right\}. \quad (43)$$

This criterion works extremely well in practice.

V. Experimental Illustration

A. Chicken Pieces Silhouettes Database

The Chicken Pieces Silhouettes Database [9] consists of 446 images of chicken pieces. Each piece belongs to one of five categories, which represent specific parts of the chicken. Each image is in binary format containing the silhouette of a particular piece. The dataset lends itself to kernelisation over standard pattern recognition via embedding shape characterizers in a uniform-dimension pattern-recognition space (difficul-

ties that e.g. edit-distance kernels naturally overcome, being able to compare silhouettes of differing size).

Pieces are placed in a natural way without considering orientation and represented by 44 pair-wise real-valued similarity measures $S_i(\omega',\omega'')$, $i=1,...,44$, derived from different parameterizations of the edit-distance kernel. We thus consider a binary class problem with 172 entities: $\Omega=\{\omega_j,\ j=1,...,N=172\}$, $y_j=\pm 1$.

B. *Secondary object features and the process of training*

We represent each entity ω_j by the N-dimensional vector of its secondary features, i.e., similarities with all the elements of the training set

$$\boldsymbol{x}_j=(x_{j1},...,x_{jN})=\big(S(\omega_j,\omega_1),...,S(\omega_j,\omega_N)\big)\in\mathbb{R}^N$$

and solve the RVM problem (8) with a large value of parameter $C>0$ and increasing values of selectivity parameter $\mu\geq 0$.

The solution of this problem is denoted as $(\hat{a}_{C\mu,1},...,\hat{a}_{C\mu,N},\hat{b}_{C\mu},\ \hat{\xi}_{C\mu,1},...,\hat{\xi}_{C\mu,N})$ defining a discriminant hyperplane in the respective N-dimensional feature space $\sum_{i=1}^{N}\hat{a}_{C\mu,i}x_i+\hat{b}_{C\mu}\gtrless 0$.

C. *Illustrative Experimental Comparison*

For each conjectural value of μ, the procedure determines the subset of *relevance secondary features* $\hat{\mathbb{J}}_{C\mu}=\{\hat{a}_{C\mu,i}\neq 0\}\subset\mathbb{J}$, namely, of the relevance training-set entities.

To provide a baseline for the proposed method of hypothetical cross-validation, we applied standard leave-one-out cross validation to each value of μ. The result is shown in Figure 1.

In particular, it is evident that the leave-one-out estimate of the generalization performance does not significantly depend on the selectivity level for this data set in practice. This implies that a standard SVM, which is obtained from the RVM formulation when $\mu=0$, is not liable to overfitting on the dataset, and further that features do not contain *complementary* information. This makes it an ideal test bed to determine the practical behavior of the non-enumerative cross-validation approach.

We thus applied the non-enumerative hypothetical cross-validation technique for the RVM, as outlined in Section IV.C. The plot of the criterion (43) is shown in Figure 2.

Note, in particular, that the selectivity setting exhibits a strong peak at $\mu\cong 500$, even in an experimental context chosen to exhibit relatively little structural risk over the majority of the tested range. Crucially, the peak does not contradict to the region of acceptable generalization performance in the leave-one-out scenario. This is indicative that the employed heuristic assumptions are sufficient to bring about a strong instantiation of selectivity parameter even in scenarios where weaker selection would be viable. Note that a strong instantiation of selectivity has potential advantages in reducing overall training times, and may be considered analogous to the rapid tending to zero of non-support object's Lagrange coefficients in an efficient implementation of the standard SVM formulation.

VI. Conclusions

In order to avoid the multiple training repetitions required by traditional cross-validation when adjusting structural-risk related parameters, we propose a novel non-enumerative hypothetical cross validation approach. In particular, we demonstrate that by making certain mild heuristic assumptions regarding the underlying distribution of the data we can derive a quantity that is analogous to VC dimensionality, albeit within a strictly information theoretic context (we show that the classical Akaike Information Criterion is a particular case).

We demonstrate the effectiveness of the non-enumerative cross-validation method on the chicken-pieces data set, and establish the method's strong instantiation tendencies with regard to the selectivity parameter of a doubly-regularized SVM variant.

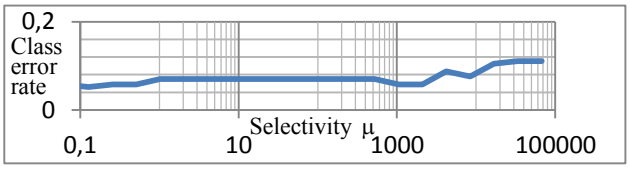

Figure 1. Leave-one-out cross validation of RVM with increasing selectivity.

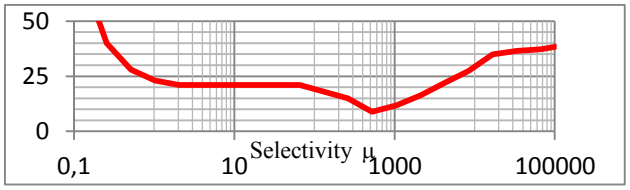

Figure 2. The criterion of hypothetical cross validation for choosing the selectivity level (43).

Acknowledgement

We would like to acknowledge support from grants of the Russian Foundation for Basic Research 11-07-00728, 13-07-13132, 14-07-00661, and from UK EPSRC Visual Media Platform grant EP/F02827X/1.

References

[1] C. Cortes, V. Vapnik. Support-Vector Networks. *Machine Learning*, 1995, 20, pp. 273-297.

[2] V. Vapnik. *Statistical Learning Theory*. John-Wiley & Sons, Inc., 1998, 736 p.

[3] A. Bab-Hadiashar, D. Suter. *Data Segmentation and Model Selection for Computer Vision: A Statistical Approach*. Springer Verlag, New York, Inc., 2000.

[4] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, Vol. 19, pp. 716-723.

[5] R. Duin, E. Pekalska, D. de Ridder. Relational discriminant analysis. *Pattern Recognition Letters*, Vol. 20, 1999, pp. 1175-1181.

[6] C. Bishop, M. Tipping. Variational Relevance Vector Machines. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 46-53. Morgan Kaufmann, 2000.

[7] L. Wang, J. Zhu, H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, Vol. 16, 2006, pp. 589-615.

[8] O. Seredin, V. Mottl, A. Tatarchuk, N. Razin, D. Windridge. Convex Support and Relevance Vector Machines for selective multimodal pattern recognition. *Proceedings of the 21th International Conference on Pattern Recognition*, Tsukuba, Japan, November 11-15, 2012. IAPR, 2012, ISSN 978-4-9906441-1-6, 2012, pp. 1647-1650.

[9] G. Andreu, A. Crespo, J.M. Valiente. Selecting the Toroidal Self-Organizing Feature Maps (TSOFM) Best Organized to Object Recognition. Proceedings of ICNN'97, vol. 2, pp. 1341-1346, Houston, Texas (USA). IEEE. June, 1997.