Theor Appl Genet (2007) 115:1003–1013
DOI 10.1007/s00122-007-0627-9

ORIGINAL PAPER

Support vector machine regression for the prediction of maize hybrid performance

S. Maenhout · B. De Baets · G. Haesaert · E. Van Bockstaele

Received: 30 May 2007 / Accepted: 2 August 2007 / Published online: 6 September 2007

Abstract Accurate prediction of the phenotypical performance of untested single-cross hybrids allows for a faster genetic progress of the breeding pool at a reduced cost. We propose a prediction method based on ε-insensitive support vector machine regression (ε-SVR). A brief overview of the theoretical background of this fairly new technique and the use of specific kernel functions based on commonly applied genetic similarity measures for dominant and co-dominant markers are presented. These different marker types can be integrated into a single regression model by means of simple kernel operations. Field trial data from the grain maize breeding programme of the private company RAGT R2n are used to assess the predictive capabilities of the proposed methodology. Prediction accuracies are compared to those of one of today's best performing prediction methods based on best linear unbiased prediction. Results on our data indicate that both methods match each other's prediction accuracies for several combinations of marker types and traits. The ε-SVR framework, however, allows for a greater flexibility in combining different kinds of predictor variables.

Communicated by M. Cooper.

S. Maenhout (✉) · G. Haesaert
Department of Plant Production, University College Ghent,
Voskenslaan 270, Gent 9000, Belgium
e-mail: Steven.Maenhout@hogent.be

B. De Baets
Department of Applied Mathematics,
Biometrics and Process Control, Ghent University,
Coupure links 653, Gent 9000, Belgium

E. Van Bockstaele
Department of Plant Production, Ghent University,
Coupure links 653, Gent 9000, Belgium

E. Van Bockstaele
ILVO, Institute for Agricultural and Fisheries Research,
Van Gansberghelaan 96, Merelbeke 9820, Belgium

Introduction

For several agronomically important plant species like maize (*Zea mays* L.), hybrid varieties constitute a considerable part, if not all, of the commercial market. Maize breeding programmes typically have a continuously evolving breeding pool at their disposal which is loosely divided into several complementary heterotic groups. New inbred lines are created by subsequent inbreeding of an initial cross or the use of doubled haploids. During their selection, these candidate lines are crossed with tester lines from a complementary heterotic group and hybrid performance is evaluated in multi-location field trials. Bernardo (1994, 1995, 1996a, b) uses linear mixed modelling to predict the performance of such an untested cross based on field trial results of related hybrids and marker data. This approach performs well considering the upper limit in prediction accuracy that is imposed by the heritability of each tested trait. Charcosset et al. (1998) show that this prediction method is superior when hybrids originate from crosses between unrelated inbred lines, which is most likely the case in commercial breeding programmes. Unfortunately, correlations between predicted and observed SCA values are too low to allow for an effective selection towards high heterosis hybrids (Bernardo 1995). Maenhout et al. (2007) demonstrate how ε-insensitive support vector machine regression (ε-SVR) can be used to screen for genetically superior inbred lines based on their

molecular marker profiles. They conclude that prior to field testing ε-SVR allows to predict the GCA component of an inbred line with adequate precision. Unfortunately this does not hold for the SCA component for which an accurate prediction method is still out of reach.

Based on these promising results as a screening tool, this paper explores the use of ε-SVR to directly predict the phenotypical performance of untested hybrids. The presented technique uses linear mixed modelling to correct unbalanced phenotypical measurements for nuisance parameters like trial, location and block effects. The corrected phenotypical values of all hybrids are used as a training set for constructing an ε-insensitive regression model in which the molecular fingerprints of each hybrid serve as predictor variables. These models can subsequently be used to predict the phenotypical values of unknown hybrids and inbred lines. The advantage of ε-SVR lies in the use of kernel functions that allow to explore nonlinear models for hybrid prediction.

Materials and methods

Data description

Phenotypical data

The phenotypical data used in this study originate from field trials that were organised as part of the grain maize breeding programme of RAGT R2n between 1998 and 2005. One hundred and five inbred lines from the Iowa stiff stalk synthetic (ISSS) heterotic group and 93 lines from the complementary Iodent group were selected on the basis of three criteria:

1. The theoretical half diallel between the 105 lines of the ISSS group and the 93 lines of the Iodent group should be as complete as possible.
2. All genetic effects should be estimable if estimation were to be done using a linear model.
3. The maximum prediction error variance (PEV) of a pairwise contrast between the random genetic components of hybrids in the selection should be minimal.

These three criteria together ensure that the selected inbred lines produce the maximum number of training samples with low PEV. The resulting data set contains 2,371 hybrids which were tested in 1,287 multi-location field trials. There are on average 34.4 hybrids tested in each of these multi-location field trials. In every location a trial is laid out as a randomised complete block design, but 67% of them include only one block, leading to an average of 1.42 replications per location. To ensure estimability, connectivity and a good model fit, the phenotypical measurements of an additional 34,140 hybrids are added to the data set as check varieties. The data set contains two levels of unbalancedness. Firstly, the 2,371 created hybrids only represent 24.3% of all possible crosses in the theoretical half-diallel. Secondly, each tested hybrid or check variety is on average present in 1.2 connected multi-location field trials demonstrating the severe unbalancedness at this secondary level. For each included plot in the data set, grain yield (q/ha at 15% moisture), grain moisture content and days until flowering were recorded. The coefficient of coancestry $\theta^{\mathrm{P}}_{ii'}$ between two inbred lines i and i' of the same heterotic group was calculated by tabular analysis from pedigree information with corrections for inbreeding and backcrossing (Emik and Terrill 1949).

Marker data

The 198 selected inbred lines were fingerprinted using co-dominant SSR and dominant AFLP markers. The 101 genotyped SSR markers are evenly distributed over the maize genome according to a proprietary linkage map of the company RAGT R2n. Due to problems identifying some SSR alleles (null alleles) only 75 markers have complete profiles over all selected inbred lines. In this study only information of these complete SSR loci was used. About 2.6% of all SSR locus/inbred line combinations was heterozygous, preventing an exact deduction of the hybrid genotype when these lines are used as parents. The average polymorphism information content (PIC) of the 75 SSR loci over 198 selected inbred lines is 0.55. The molecular coefficient of coancestry $\theta^{\mathrm{S}}_{ii'}$ between two inbred lines i and i' of the same heterotic group was calculated from SSR data as described by Bernardo (1993),

$$\theta^{S}_{ii'} = \frac{P_{ii'} - \frac{1}{2}(\bar{P}_{i.} + \bar{P}_{i'.})}{1 - \frac{1}{2}(\bar{P}_{i.} + \bar{P}_{i'.})}, \qquad (1)$$

where $P_{ii'}$ is the average allele identity over all SSR marker loci between inbred i and i' defined as

$$P_{ii'} = \frac{1}{4s}\sum_{k=1}^{s} I(i_{k_{\mathrm{m}}}, i'_{k_{\mathrm{m}}}) + I(i_{k_{\mathrm{m}}}, i'_{k_{\mathrm{p}}}) + I(i_{k_{\mathrm{p}}}, i'_{k_{\mathrm{m}}}) + I(i_{k_{\mathrm{p}}}, i'_{k_{\mathrm{p}}}),$$

where $i_{k_{\mathrm{m}}}$ represents the maternal allele of inbred line i for locus k, while $i_{k_{\mathrm{p}}}$ represents the paternal allele. $I(i_{k_{\mathrm{m}}}, i'_{k_{\mathrm{m}}})$ returns one if the maternal allele on locus k of individual i is equal to the maternal allele of that same locus of individual i' and 0 otherwise. $\bar{P}_{i.}$ represents the average allele identity between inbred line i and all inbred lines of the complementary heterotic group. This formulation allows for incomplete homozygosity of inbred lines. $\bar{P}_{i.}$ represents

the average allele identity between inbred line i and all lines of its complementary heterotic group.

AFLP data is generated according to the protocol of Vos et al. (1995) using 11 PstI–MseI (P12/M47, P13/M47, P12/M59, P13/M48, P12/M61, P13/M49, P12/M62, P13/M59, P12/M50, P12/M48, P12/M49) and four EcoRI–MseI primer combinations (E38/M51, E39/M55, E39/M59, E46/M59) (Vuylsteke et al. 1999). The EcoRI and MseI primers each had three selective nucleotides, while there were only two for the PstI primers. There was preference for the PstI–MseI primer combinations as the resulting markers are likely to be more evenly distributed over the maize genome than EcoRI–MseI markers (Vuylsteke et al. 1999; Castiglioni et al. 1999). These 15 primer combinations produced 569 polymorphic bands for the 198 selected inbred lines. To calculate the molecular coefficient of coancestry $\theta^A_{ii'}$ between two inbred lines i and i' of the same heterotic group based on dominant AFLP marker data, Eq. (1) was used but the proportion of shared AFLP alleles $P_{ii'}$ was calculated according to the Jaccard similarity measure as

$$P_{ii'} = \frac{a_{ii'}}{a_{ii'} + b_{ii'} + c_{ii'}}, \tag{2}$$

where $a_{ii'}$ represents the number of bands common to both individuals i and i' while $b_{ii'}$ represents the number of bands unique to i and $c_{ii'}$ those unique to i'.

Data analysis

Linear mixed model

As the data suffers from severe unbalancedness, a linear mixed model is the recommended approach for correcting the phenotypical measurements for nuisance factors like trial, location and block effects. The used model is quite similar to that proposed by Bernardo (1994) but the actual plot measurements are used instead of averages over locations and blocks:

$$\mathbf{y} = \mu + \mathbf{X}_t\mathbf{t} + \mathbf{X}_l\mathbf{l} + \mathbf{X}_b\mathbf{b} + \mathbf{Z}_c\mathbf{c} + \mathbf{Z}_I\mathbf{a}_I + \mathbf{Z}_O\mathbf{a}_O + \mathbf{Z}_d\mathbf{d} + \mathbf{e}. \tag{3}$$

$\mathbf{y}$ represents a vector containing the trait responses for each plot in the data set and μ represents the global phenotypical mean. $\mathbf{t}$ is a vector containing the fixed multi-location trial effects, $\mathbf{l}$ contains the fixed effects for each location nested within a multi-location trial and $\mathbf{b}$ represents the fixed block effects, nested within each location. Vector $\mathbf{c}$ contains the random genotypical effects for all checks. $\mathbf{a}_I$ and $\mathbf{a}_O$ are vectors containing GCA effects for the inbred lines belonging to the ISSS and Iodent heterotic groups, respectively, while $\mathbf{d}$ contains the SCA effects for each of the 2,371 hybrids. $\mathbf{e}$ contains a random residual error for each plot in the data set. The coincidence matrices $\mathbf{X}_t$, $\mathbf{X}_l$, $\mathbf{X}_b$, $\mathbf{Z}_c$, $\mathbf{Z}_I$, $\mathbf{Z}_O$ and $\mathbf{Z}_d$ link each entry in the $\mathbf{y}$ vector with the appropriate effect. The levels of the nested trial, location and block effects are sometimes confounded in which case the higher level effect is set to 0. No explicit GxE terms or heterogeneous residual variances were fitted into Eq. (3). The expected improvements of these more elaborate models could not be verified because of computational limitations caused by the size of the data set and its severe unbalancedness. Furthermore, these models are not handled by the benchmark method described by Bernardo (1994, 1995, 1996a, b) and would therefore exclude an objective comparison.

The covariance matrix $\mathbf{G}$ for the random effects in the model can be represented as

$$\mathbf{G} = \begin{bmatrix} \mathbf{I}\sigma^2_c & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_I\sigma^2_I & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_O\sigma^2_O & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D}\sigma^2_d & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_r \end{bmatrix} \tag{4}$$

The matrices $\mathbf{A}_I$ and $\mathbf{A}_O$ model the covariance between inbred lines of the ISSS and Iodent heterotic group, respectively. Usually the covariance between two hybrids h_{ij} and $h_{i'j'}$, where lines i and i' belong to the ISSS group and lines j and j' belong to the Iodent group, is modelled as (Stuber and Cockerham 1966)

$$\mathrm{Cov}(h_{ij}, h_{i'j'}) = \theta_{ii'}\sigma^2_I + \theta_{jj'}\sigma^2_O + \theta_{ii'}\theta_{jj'}\sigma^2_d, \tag{5}$$

where $\theta_{ii'}$ is the coefficient of coancestry between two inbred lines i and i' of the ISSS heterotic group and $\theta_{jj'}$ between two inbred lines j and j' of the Iodent group. The coefficient of coancestry can be calculated based on pedigree information (θ^{P}), but also from SSR (θ^{S}) or AFLP data (θ^{A}). The three components of Eq. (5) allow to construct the matrices $\mathbf{A}_I$, $\mathbf{A}_O$ and $\mathbf{D}$ using the described coefficients of coancestry. These alternative formulations are compared by means of the restricted log-likelihood of the model given the data, keeping the fixed effects structure constant. The covariance matrix for the checks is assumed to be an identity matrix. If pedigree or marker data were available for these checks, including this information in the covariance matrix $\mathbf{G}$ would improve the model fit as proposed by Bernardo (1995). The variance parameters σ^2_c, σ^2_I, σ^2_O, σ^2_d and σ^2_r are estimated through REML optimisation by means of the average information algorithm as implemented in the software tool ASReml (Gilmour et al. 2002).

The phenotypical value of each hybrid is estimated as the average of its measurements in the data set, albeit with correction for trial, location and block effects. The vector

of corrected phenotypical values is therefore obtained as (Bernardo 1994, 1995, 1996a, b)

$$\hat{\mathbf{y}}_c = (\mathbf{Z}_d'\mathbf{Z}_d)^{-1}\mathbf{Z}_d'(\mathbf{y} - \mu - \mathbf{X}_t\mathbf{t} - \mathbf{X}_l\mathbf{l} - \mathbf{X}_b\mathbf{b}). \quad (6)$$

The elements of $\hat{\mathbf{y}}_c$ are used as a training set for building the ε-SVR prediction model. Apart from training an ε-SVR model we also implemented the prediction system proposed by Bernardo (1994, 1995, 1996a, b). A validation subset $\hat{\mathbf{y}}_{cv}$ of size l' is predicted from the remaining entries $\hat{\mathbf{y}}_{ct}$ as

$$\hat{\mathbf{y}}_{cv} = \mathbf{C}_{vt}\mathbf{V}_{\mathrm{t}}^{-1}\hat{\mathbf{y}}_{ct}. \quad (7)$$

$\mathbf{C}_{vt}$ is an $l' \times (l\text{–}l')$ matrix containing the covariances between validation and training hybrids. $\mathbf{V}_{\mathrm{t}}$ is the variance–covariance matrix of the l–l' training hybrids. Elements of $\mathbf{C}_{vt}$ and non-diagonal elements of $\mathbf{V}_t$ are computed using Eq. (5). The ith diagonal element of $\mathbf{V}_{\mathrm{t}}$ is equal to $\sigma_I^2 + \sigma_O^2 + \sigma_d^2 + \frac{\sigma_r^2}{n_i}$, where n_i is the number of records of the ith hybrid in the training set. The prediction accuracy of Bernardo's method is established using a leave-one-out cross-validation. This means that each of the 2,371 hybrids are individually predicted using a vector $\hat{\mathbf{y}}_{ct}$ containing the corrected phenotypical effects of the 2,370 remaining hybrids. The algorithm was implemented in C++ using the matrix routines provided in the GNU Scientific Library (Galassi et al. 1998).

ε-insensitive support vector machine regression

Support vector machines (SVM) are a set of unsupervised learning methods developed by Vapnik (1995) for classification and regression. A good tutorial on SVM classification is given by Burges (1998), while Smola and Schölkopf (2004) present the underlying ideas of support vector machines for regression (SVR). In an SVR setting we represent each training sample i as a couple consisting of a vector $\mathbf{x}_i \in \mathcal{X}$ and a scalar $y_i \in \mathbb{R}$. If we want to learn the phenotypical performance of a hybrid maize plant based on the molecular fingerprints of its two parental inbred lines, we could consider $\mathcal{X}$ as a binary space of n dimensions where n is the total number of possible alleles that make up a molecular fingerprint. For each hybrid i, the entries in the vector $\mathbf{x}_i$ are set to one if one of the homozygous parents carries the corresponding allele or –1 otherwise. y_i then equals the phenotypical response of hybrid i for the trait under study.

In ε-insensitive support vector machine regression (ε-SVR) the goal is to find a function $f(\mathbf{x})$ that deviates at most ε from the target value y for each training sample $0 < i \leq l$ in the data set. Initially we restrict the possible set of solutions to linear functions like

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \qquad \text{where } \mathbf{w} \in \mathbb{R}^n \text{ and } b \in \mathbb{R},$$

but sometimes several linear solutions to this problem might exist. We therefore include the additional constraint that the norm of the weight vector $\mathbf{w}$ should be as small as possible. This last condition generates simple (flat) solutions which avoid overfitting the training data. Figure 1 depicts a regression problem for which no linear solution exists for the given width ε of the insensitivity tube. Each training sample i is therefore allowed to have a slack variable $\zeta_i = y_i - f(\mathbf{x}_i) - \varepsilon$ in case $f(\mathbf{x}_i)$ underestimates y_i or $\zeta_i^* = f(\mathbf{x}_i) - y_i - \varepsilon$ in case of overestimation. These training errors should obviously be minimised together with the Euclidean norm of the weight vector $\mathbf{w}$ which allows for the formulation

$$\text{Minimise } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}(\zeta_i + \zeta_i^*)$$
$$\text{subject to} \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b & \leq \varepsilon + \zeta_i \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i & \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* & \geq 0 \end{cases} \quad (8)$$

The constant $C > 0$ determines the trade-off between the flatness of f and the extent to which deviations larger than ε are tolerated. The parameters C and ε are problem-dependent but can for example be determined by means of a simple grid search in combination with some cross-validation routine or more elaborate strategies like gradient descent methods (Chapelle et al. 2002). The inequality constraints are included into the minimisation problem through the use of Lagrange multipliers which allows for the primal formulation:

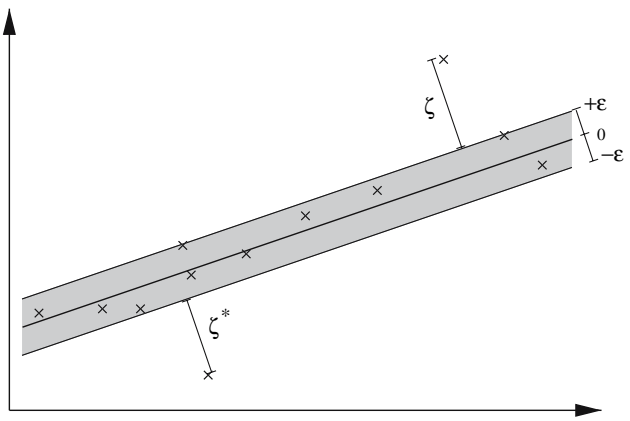

Fig. 1 A one-dimensional linear function $f(\mathbf{x})$ where all but two training samples lie within the ε-SVR insensitivity tube of width 2ε

$$L := \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}(\zeta_i + \zeta_i^*) - \sum_{i=1}^{l}(\eta_i\zeta_i + \eta_i^*\zeta_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \zeta_i - y_i + \langle \mathbf{w}, \mathbf{x}_i\rangle + b) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \zeta_i^* + y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle - b) \quad (9)$$

The partial derivatives of L with respect to the unknown function parameters b, $\mathbf{w}$, ζ_i and ζ_i^* should become 0 at the optimal point:

$$\partial_b L = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \quad (10)$$

$$\partial_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\mathbf{x}_i = 0 \quad (11)$$

$$\partial_{\zeta_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \quad (12)$$

Substituting Eqs. (10)–(12) into Eq. (9) allows for the dual formulation

$$\text{maximise}\begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i, \mathbf{x}_j\rangle \\ \quad -\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{cases}$$
$$\text{subject to}\ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \ \text{ and } \ \alpha_i, \alpha_i^* \in [0, C], \quad (13)$$

and allows us to rewrite f as:

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle \mathbf{x}_i, \mathbf{x}\rangle + b. \quad (14)$$

Eq. (14) shows that f is specified as a linear combination of all training samples $\mathbf{x}_i$ expressed as dot products. The Karush-Kuhn-Tucker (KKT) conditions state that at the solution of the maximisation problem of Eq. (13) the product between the dual variables $\alpha_i^{(*)}$ and their corresponding inequality constraints of Eq. (8) becomes 0. This basically means that only when $|f(\mathbf{x}_i)–\mathbf{y}_i| \geq \varepsilon$ the coefficients α_i or α_i^* can be non-zero. All samples inside the ε-tube are therefore not used in the formulation of f. All other training samples with nonvanishing coefficients $\alpha_i^{(*)}$ are called the support vectors, hence the name support vector machines.

If we preprocess the training samples $\mathbf{x}_i$ by a map $\phi : \mathcal{X} \to \mathcal{F}$ into a higher dimensional space named the feature space $\mathcal{F}$ and solve the linear regression there, we can state Eq. (14) as

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle \phi(\mathbf{x}_i), \phi(\mathbf{x})\rangle + b. \quad (15)$$

Depending on the map ϕ, this approach effectively allows us to create non-linear functions f. When predicting y for an unknown example $\mathbf{x}$ using the in feature space learned linear function f, Eq. (15) obliges us to apply the mapping ϕ to this new case as well as to all training samples and subsequently make the dot product between them. This approach is often not computationally feasible, so we use instead a symmetric kernel function $k(\mathbf{x}_i,\mathbf{x}) = \langle \phi(\mathbf{x}_i),\phi(\mathbf{x}) \rangle$ that gives us directly the dot product in feature space. This shortcut allows us to reformulate Eq. (15) as

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(\mathbf{x}_i, \mathbf{x}) + b. \quad (16)$$

Not all symmetric functions over $\mathcal{X} \times \mathcal{X}$ are kernels that can be used in an SVM. Since a kernel function k is related to an inner product it has to satisfy some conditions that arise naturally from the definition of an inner product and are given by Mercer's theorem: the kernel function has to be positive semi-definite (PSD). A commonly used kernel function is the Gaussian kernel defined as

$$k(\mathbf{x}, \mathbf{z}) = \exp(-\gamma\|\mathbf{x} - \mathbf{z}\|^2), \quad (17)$$

where γ is a kernel specific parameter which allows to find a linear function in an infinitely large feature space (Shawe-Taylor and Cristianini 2004). Most all-round kernels like the Gaussian or polynomial kernel require the knowledge of one or several additional kernel parameters. The use of context specific kernel functions, however, can avoid the computationally exhausting grid searches needed to identify these parameter values that allow a minimal generalisation error.

Dot products in feature space are in fact measures of similarity between cases, so the use of PSD genetic similarity measures as kernel functions is a valid option. The Jaccard similarity measure of Eq. (2) is commonly used when genotyping is based on dominant molecular markers like AFLP. As this similarity measure is PSD (Gower and Legendre 1986), it can be used as a kernel function in an ε-SVR.

A useful PSD genetic similarity measure for co-dominant markers is the complement of the Modified Rogers' distance (Wright 1978; Goodman and Stuber 1983) or MRD defined as

$$s_{kl}^W = 1 - d_{kl}^W \quad \text{with} \quad d_{kl}^W = \frac{1}{\sqrt{2s}}\sqrt{\sum_{i=1}^{s}\sum_{j=1}^{n_i}(p_{ij}^k - p_{ij}^l)^2},$$

where s is the number of genotyped loci, n_i is the number of alleles for locus i and p_{ij}^k, p_{ij}^l represent the allele

frequency for the jth allele of locus i for individual k and l, respectively. As demonstrated in Melchinger (1999) there is a linear relationship between ΔH, the panmictic-mid-parent heterosis and $(d_{kl}^{W})^2$ under the assumption of biallelism and absence of epistasis. Therefore, this similarity measure should prove itself useful when used as a kernel function for hybrid prediction.

The weighted sum of two PSD matrices produces a PSD matrix as long as the weights are positive. Computing the weighted sum of different kernel functions therefore creates a new kernel function. Returning to the concept of a feature space this operation has the effect of augmenting the dimensions of the feature space related to the first kernel, with the dimensions of the feature space related to the second kernel (Shawe-Taylor and Cristianini 2004). When we apply this to the Jaccard and MRD kernel functions we have a way to combine SSR and AFLP data into a single regression function. We call the resulting function the Jaccard–MRD kernel.

Cross-validation and grid search

To assess the generalisation error of an ε-SVR model we rely on a leave-one-out cross-validation procedure. It is, however, infeasible to redo the REML optimisation for each reduced training set as removing records of a randomly chosen hybrid for cross-validation would cause connectivity issues in the linear mixed model. These problems include inestimable fixed nuisance parameters and biased estimations of variance components. It is therefore assumed that differences between the estimators of the fixed nuisance parameters, calculated using only the data from the training hybrids, and those estimated using all data, are negligible. The reported results therefore do not account for the loss of prediction accuracy in the linear mixed model caused by data reduction.

For each hybrid in the data set, a different ε-SVR model is trained using the corrected phenotypical values $(\hat{\mathbf{y}}_c)$ of the remaining hybrids as a training set. When $\hat{\mathbf{y}}_c$ of a trait is used as predictand, the square root of the broad-sense heritability of that trait upper bounds the correlation between the true and predicted phenotypical values as explained by Bernardo (1996a). These broad-sense heritability estimates should not be compared to the usual narrow-sense heritability estimates calculated on an entry-mean basis.

Building an ε-regression model from training values requires values for ε, C and γ when using the Gaussian kernel. Finetuning these variables can greatly improve the generalisation capacity of the prediction system. To find the optimal values a grid search was performed as described by Hsu et al. (2003). During this grid search all combinations of ε, C and if necessary γ were tested for each cross-validation routine, where ε and γ ranged from 2^{-15} to 2^{4} and C ranged from 2^{-5} to 2^{15}. The software libSVM (Chang and Lin 2001), which allows easy integration of non-standard kernel functions, was used for all regressions. Calculations were performed on a Linux cluster containing eight nodes, each having two Dual-Core Intel® Xeon® CPU 3.00 GHz processors, 1 Gb of RAM and running a 2.6.5 kernel.

When reporting prediction accuracies, several artificial measures could be used to compare models and techniques. A commonly used measure of prediction accuracy is the standard error defined as the root of the summed squared differences between the actual and the predicted values divided by the number of predictions. Although this measure allows for easy comparison between different models and data sets, it is dependent on the unit of measurement of the response variable. Comparing accuracies of similar techniques or models on traits measured in a different unit or scale is therefore not possible. Interpreting standard errors is also quite hard when the reader has no reference for comparing the obtained results. Another commonly used measure is the Pearson correlation ρ between the actual and the predicted value. This correlation, expressed as a number between 0 and 1, is however dependent on the variance of the predictor variable and resulting predictions. The larger this variance, the larger the obtained correlation will be. This means for example that the correlation between the actual and predicted value for a regression on yield will be larger in natural populations compared to advanced breeding pools with lower yield variance. This property makes it hard to compare published results between prediction methods when different data sets are used. As both criteria seem to cover each others' weaknesses we compare the different prediction systems by calculating the Pearson correlation as well as the standard error.

Results

Linear mixed model fit

The average coefficient of coancestry calculated from pedigree data differs substantially from the averages calculated from SSR or AFLP data as can be seen from Table 1. Despite the apparent differences between the mean values for θ^{P}, θ^{A} and θ^{S}, the Spearman rank correlations between these estimators are moderately high. The AFLP-based coefficients seem to represent an intermediate value between the high SSR- and low pedigree-based coefficients. As can be seen from Table 2, the observed correlations between the two marker-based coefficients of

Table 1 Minimum, maximum and average coancestries based on pedigree (θ^P), AFLP (θ^A) and SSR (θ^S) for the two heterotic groups used in this study

	θ^P	θ^A	θ^S
Iodent			
Average	0.27	0.38	0.45
Minimun	0	0.04	0.01
Maximum	0.88	0.99	0.98
ISSS			
Average	0.17	0.23	0.31
Minimum	0	0	0
Maximum	0.78	0.94	0.95

coancestry are higher than the correlations between a marker-based and a pedigree-based coefficient for both heterotic groups. However, AFLP-based estimators are closer to the pedigree-based coefficients than the SSR-based alternatives. Apparently all calculated correlations within the Iodent group are greater than those of the ISSS group.

We use the log-likelihood resulting from the REML optimisation process to determine the best fitting covariance structure for Eq. (4). Table 3 gives an overview of these log-likelihoods for the linear mixed model of Eq. (3) where the matrices $\mathbf{A}_I$, $\mathbf{A}_O$ and **D** are either considered diagonal or constructed according to Eq. (5) using pedigree, SSR or AFLP data for the calculation of the coefficients of coancestry. The models with a non-diagonal covariance matrix for the SCA values always have a lower log-likelihood than their diagonal counterparts. This means that the covariance between SCA values should be modelled as 0 as it seems to fit better than the product of both coefficients of coancestry as in Eq. (5).

The model with AFLP-based **A** matrices and an identity **D** matrix results in the highest log-likelihood for all traits under study. A model with SSR-based **A** matrices and a diagonal **D** matrix gives the second highest log-likelihood for yield, but performs worse than the pedigree-based equivalent for moisture content and days until flowering. These results indicate that the AFLP-based coefficient of coancestry approximates better the actual relatedness between hybrids compared to the pedigree-based and even the SSR-based coefficient for this data set. All subsequent regressions and predictions are therefore based on the results of the linear mixed models with AFLP-based **A** and diagonal **D** matrices in Eq. (4).

Table 2 Spearman rank correlations between coefficients of coancestry based on pedigree (θ^P), AFLP (θ^A) and SSR (θ^S) data for the Iodent and ISSS heterotic groups

ρ	Iodent	ISSS
$\theta^P \leftrightarrow \theta^A$	0.79	0.69
$\theta^P \leftrightarrow \theta^S$	0.75	0.67
$\theta^A \leftrightarrow \theta^S$	0.90	0.77

Table 3 Restricted log-likelihoods for the linear mixed model of Eq. (3) with fixed nuisance factors but different formulations for **G**. The covariance matrices for GCA and SCA effects are either diagonal, based on pedigree, SSR or AFLP data

A	**D**	Yield	Moisture (%)	Flowering
Diagonal	Diagonal	–588609	–201915	–158515
Pedigree	Diagonal	–588590	–201872	–158498
Pedigree	Pedigree	–588659	–202056	–158571
SSR	Diagonal	–588585	–201879	–158504
SSR	SSR	–588681	–202142	–158600
AFLP	Diagonal	**–588583**	**–201855**	**–158487**
AFLP	AFLP	–588639	–201961	–158537

ε-SVR

When testing new hybrid prediction algorithms, the main interest lies in the estimation of the total genetic value of untested hybrids. We use the corrected phenotypical values in vector $\hat{\mathbf{y}}_c$ from Eq. (6) as a training set for building a regression model. By means of the standard leave-one-out cross-validation strategy the predictive capabilities of the different kernels are compared to each other. Table 4 gives an overview of the obtained correlations and standard deviations for the different combinations of trait, marker type and kernel functions. The last column represents the leave-one-out cross-validation accuracy of the prediction by means of Eq. (7) using marker-based coefficients of coancestry to model $\mathbf{C}^{vt}$ and $\mathbf{V}^t$.

When the molecular information is restricted to micro satellite data, the ε-SVR based models, albeit with a minimal difference, provide better prediction accuracies than Bernardo's method. Comparing the three kernel functions, we notice that the two nonlinear kernel functions always perform slightly better than the linear one. This observation demonstrates the advantage of performing a linear regression in a kernel induced feature space. The similarity based MRD kernel function performs just as good as the Gaussian kernel but does not require the finetuning of an additional kernel parameter so using MRD to build an optimised prediction model takes far less computation time. Prediction accuracies of ε-SVR and Bernardo's method are also very similar when the molecular fingerprints of the inbred lines are restricted to AFLP markers. For yield and days until flowering ε-SVR is slightly superior, while Bernardo's method is preferred for moisture content. Again the nonlinear kernels perform better than their linear counterpart

Table 4 Standard leave-one-out prediction accuracies, expressed as Pearson correlations and standard errors (between brackets), on corrected phenotypical values for yield, moisture content and days until flowering

	Linear	Gaussian	MRD	Bernardo
SSR				
Yield	0.56 (6.8)	0.58 (6.67)	0.58 (6.67)	0.57 (6.72)
Moisture content	0.83 (1.19)	0.84 (1.16)	0.84 (1.14)	0.84 (1.16)
Flowering	0.62 (1.18)	0.63 (1.16)	0.63 (1.16)	0.62 (1.18)
	Linear	Gaussian	Jaccard	Bernardo
AFLP				
Yield	0.56 (6.78)	0.58 (6.64)	0.57 (6.75)	0.57 (6.72)
Moisture content	0.83 (1.18)	0.84 (1.14)	0.84 (1.16)	**0.84** (**1.13**)
Flowering	0.61 (1.18)	0.63 (1.16)	0.62 (1.18)	0.62 (1.17)
	Linear	Gaussian	Jaccard–MRD	Bernardo
AFLP + SSR				
Yield	0.56 (6.8)	**0.58** (**6.63**)	0.57 (6.7)	–
Moisture content	0.83 (1.18)	0.84 (1.14)	0.84 (1.14)	–
Flowering	0.61 (1.19)	**0.63** (**1.15**)	0.63 (1.16)	–

The results are presented according to the type of features (SSR, AFLP or both) and the type of kernel function used during the analysis. The last column represents the accuracy of the predictions obtained with Bernardo's method (Bernardo 1994, 1995, 1996a, b). The prediction method with the highest correlation and lowest standard error is typesetted in bold for each trait

and the parameter free Jaccard-based kernel function provides a valid alternative to the Gaussian kernel.

When we need to decide between SSR- and AFLP-based features, we notice that for each trait under study, the AFLP markers provide equal or slightly better prediction accuracies than the SSR markers. Examining the restricted log-likelihood of the linear mixed model revealed the same preference for the dominant AFLP marker data. In either case the differences are minimal to say the least so these conclusions should not be generalised to other data sets. For yield and days until flowering combining the information of SSR and AFLP markers provides the highest prediction accuracy over all applied methods but the gain in precision is minimal as both sets of markers seem to be equally informative in this case.

The maximum obtained Pearson correlations using an ε-SVR based model are 0.58, 0.84 and 0.63 for yield, moisture content and days until flowering, respectively, while these are 0.57, 0.84 and 0.62 for Bernardo's method. We can therefore conclude that ε-SVR predictions are at least as accurate as the corresponding analyses using Bernardo's method. The maximum correlations, calculated as the square root of the trait's broad sense heritability, are 0.66, 0.87 and 0.68 for yield, moisture content and days until flowering, respectively. It should be clear that both frameworks predict close to the theoretical maximum for moisture content and that there is still some room for improvement in days until flowering and especially yield prediction. The ε-SVR framework should allow for these traits to tighten the remaining gap between the theoretical and obtained correlations for example by means of feature selection methods. The gradient descent based R2W2 technique described by Weston et al. (2000) and the greedy recursive feature elimination (RFE) described by Guyon et al. (2002) are examples of such methods that allow for the identification of markers that have little or no contribution to the prediction model. Besides the advantage of identifying key markers which could be used as a starting point for more detailed association studies, it is to be expected that removing the useless features shall improve the obtained prediction accuracies; however, further study is required to ascertain this point. Another possible road to improvement is to design specific kernel functions for hybrid prediction. This allows to encode prior knowledge of the learning task into the feature space in which the regression takes place. The advantages of engineering a case-specific kernel function are exemplified by Zien et al. (2000) who designed a kernel for the identification of translation initiation sites in DNA code which resulted in a significantly improved recognition performance compared to the standard kernel functions.

Discussion

Bernardo's method is currently one of the best known methods for the prediction of the phenotypical performance of maize hybrids originating from crosses between unrelated lines, as is the case for most of today's commercial hybrids. We evaluated the use of ε-insensitive support vector machine regression, as an alternative to Bernardo's method, on a real maize breeding data set from the private breeding company RAGT R2n. Maenhout et al. (2007) applied ε-SVR as a screening tool for the genetic components of newly created inbred lines. The idea is now to train the ε-SVR algorithm to directly predict the phenotypical values of maize hybrids based on the molecular marker scores of both parental inbred lines and compare the obtained prediction accuracies with those of Bernardo's method. The field trial data resulting from a commercial breeding programme are typically very unbalanced and therefore linear mixed modelling is used to adjust the phenotypical measures for location, trial and block effects. For each hybrid, the average of the corrected plot measurements for yield, grain moisture contents and days until flowering are used as predictands while the AFLP- and

SSR-based molecular fingerprints of the parental inbred lines serve as predictor variables.

We calculated the coefficients of coancestry based on pedigree, SSR and AFLP data for all pairwise combinations of inbred lines within each of the two heterotic groups. The Spearman rank correlations between the obtained similarity measures are moderately high but the marker-based coefficients generally indicate a higher level of relatedness between the individual lines. This discrepancy might be explained by the unequal parental contributions that can occur after several generations of inbreeding during line development. A standard pedigree analysis is not able to detect these shifts and assumes equal contributions from both parents. Another possible cause of bias is the assumption of unrelated ancestor individuals which is often impossible to verify. As the described deviations are higher for the ISSS lines we can assume that these departures from theoretical assumptions are more pronounced within this heterotic group.

The resulting log-likelihood of the REML procedure for estimating the variance components of the linear mixed model allows to identify the best fitting covariance structure. For the data set at hand, the likelihood of the model with a diagonal covariance matrix **D** for the SCA effects is higher than the pedigree-, AFLP- and SSR-based alternatives. This result has also been observed in other data sets (Piepho H.P., 2006 personal communication at the session "BLUP in Plant Breeding", XIII EUCARPIA Biometrics in Plant Breeding Section Meeting, Zagreb, Croatia) and demonstrates that the base assumptions underlying the derivation of Eq. (5) in Stuber and Cockerham (1966), in particular the absence of linkage disequilibrium and different effects of the same alleles in the two populations, do not hold in an advanced breeding pool. Moreover Eq. (5) is a simplification, leaving out all interaction terms besides the dominance effect and therefore assuming that epistasis is negligible. We also noticed that using products of coefficients of coancestry as entries in **D** does not guarantee a positive definite covariance matrix for the SCA values which is counterintuitive and can lead to convergence problems of the REML algorithm.

The AFLP-based coefficient of coancestry is preferred when modelling the covariance between the GCA effects of the parental inbred lines although the likelihood of a model using SSR-based coancestries is comparable. This observation exemplifies the superiority of marker-based coefficients of coancestry over theoretical pedigree-based values. Marker similarities are corrected for the difference between identity in state and identity by descent by means of the average marker similarity of each inbred line with all inbred lines of the complementary heterotic group. As indicated by Bernardo et al. (1996) this approach assumes homogeneous allele frequencies among these heterotic groups. As this was generally not the case for the Iodent and ISSS group in this study, the presented coefficients of coancestry are biased. It is to be expected that the model fit will improve when the marker-based coefficients of coancestry are derived from estimators of parental contribution as described in Bernardo et al. (2000). Unfortunately, the elaborate pedigree of the 198 selected inbred lines does not allow the fingerprinting of all ancestral individuals to calculate these parental contributions from SSR or AFLP similarities. This will generally be the case when working with historically evolved heterotic groups.

By using the most likely linear mixed model we can correct the phenotypical values for each hybrid for nuisance factors and use these estimators as a training set for the construction of an ε-SVR model. Correlations between real and predicted phenotypical values by means of a leave-one-out cross-validation show that the non-linear kernels perform better than their linear counterpart for every combination of trait and marker type. This demonstrates the advantage of performing a linear regression in a kernel induced feature space. These non-linear kernels generally allow to match or slightly improve the accuracy of the currently best performing prediction method for crosses between unrelated inbred lines. The training of an ε-SVR model does, however, assume the knowledge of several parameters like the width ε of the insensitivity tube, the error weighting variable C and possibly one or several kernel function parameters. These parameters can be optimised by a simple grid search in combination with cross-validation routines but this can become computationally exhausting when the number of required kernel parameters is large. Subject-specific kernel functions like the presented Jaccard measure, MRD and their linear combinations can avoid the necessity of extra kernel parameters while allowing similar prediction accuracies. Both the Jaccard measure and the complement of the modified Rogers' distance are PSD similarity measures and therefore represent a dot product in some feature space. In practice, the requirement of a kernel function to be PSD turns out to be a very strict assumption. Several references can be found where a symmetric non-PSD similarity function is used within the standard SVM framework as a heuristic approach (Bahlmann et al. 2002; Decoste and Schölkopf 2002; Haasdonk and Keysers 2002). Problems like non-convexity of the optimisation problem can be handled by adding an additional term to the objective function of Eq. (14) as described in Fan et al. (2005). This approach guarantees that the optimisation process converges to a stationary point but only in the case of a PSD kernel function this point is the unique optimal value. This information leads one to suspect that several other similarity measures, PSD or not, and their linear combinations

could increase the prediction accuracy of ε-SVR models but further study is obviously needed to ascertain this. Another advantage of the ε-SVR methodology is the easy integration of different types of molecular and even descriptive morphological data as features. As there is no straightforward way to incorporate all this information into the covariance matrices of Bernardo's method, ε-SVR allows for a greater flexibility when the prediction system has to be implemented into an existing breeding programme. Easy feature selection heuristics like the greedy recursive feature elimination (Guyon et al. 2002) should allow for the identification of specific molecular markers and possibly parental morphological properties that are crucial for the construction of the prediction model. When evaluating new inbred lines one can make the trade-off between the cost of collecting a certain feature and the increase in prediction accuracy that this feature represents.

To conclude we can state that, although further comparisons using other data sets are necessary, the presented ε-SVR models can generally compete with Bernardo's method. Parameter optimisation, feature selection algorithms and problem-specific kernel functions are several promising aspects of this recent technique which need further investigation in the context of hybrid prediction.

Acknowledgments The authors would like to thank the people from RAGT R2n for their unreserved and open minded scientific contribution to this research. We are also very grateful to Stijn Vansteelandt, Jan De Riek and Peter Dawyndt for discussions on linear mixed modelling, genotyping by means of AFLP markers and cluster computing.

References

Bahlmann C, Haasdonk B, Burkhardt H (2002) On-line handwriting recognition with support vector machines – a kernel approach. In: Proceedings of the 8th international workshop on frontiers in handwriting recognition. IEEE Computer Society, Washington, pp 49–54

Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet 85:1055–1062

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20–25

Bernardo R (1995) Genetic models for predicting maize single-cross performance in unbalanced yield trial data. Crop Sci 35:141–147

Bernardo R (1996a) Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Sci 36:50–56

Bernardo R (1996b) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:872–876

Bernardo R, Murigneux A, Karaman Z (1996) Marker-based estimates of identity by descent and alikeness in state among maize inbreds. Theor Appl Genet 93:262–267

Bernardo R, Romero-Severson J, Ziegle J, Hauser J, Joe L, Hookstra G, Doerge R (2000) Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP and SSR data. Theor Appl Genet 100:552–556

Burges C (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167

Castiglioni P, Ajmone-Marsan P, van Wijk R, Motto M (1999) AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. Theor Appl Genet 99:425–431

Chang C, Lin C (2001) LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsv cited 20 December 2006

Chapelle O, Vapnik V, Bousquet O, Mukherjee S (2002) Choosing multiple parameters for support vector machines. Mach Learn 46:131–159

Charcosset A, Bonnisseau B, Touchebeuf O, Burstin J, Dubreuil P, Barriére Y, Gallais A, Denis JB (1998) Prediction of maize hybrid silage performance using marker data: comparison of several models for specific combining ability. Crop Sci 38:38–44

Decoste D, Schölkopf B (2002) Training invariant support vector machines. Mach Learn 46:161–190

Emik L, Terrill C (1949) Systematic procedures for calculating inbreeding coefficients. J Hered 40:51–55

Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training SVM. J Mach Learn Res 6:1889–1918

Galassi M, Davies J, Theiler J, Gough B, Priedhorsky R, Jungman G, Booth M (1998) GNU scientific library reference manual, 2nd edn. Available via http://www.gnu.org/software/gsl cited 20 December 2006

Gilmour A, Gogel B, Cullis B, Welham S, Thompson R (2002) ASREML user guide release 1.0. VSN International Ltd.

Goodman M, Stuber C (1983) Races of maize: VI. Isozyme variation among races of maize in Bolivia. Maydica 28:169–187

Gower J, Legendre P (1986) Metric and euclidean properties of dissimilarity coefficients. J Class 3:5–48

Guyon I, Weston J, Barnhil S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422

Haasdonk B, Keysers D (2002) Tangent distance kernels for support vector machines. In: Proceedings of the 16th international conference on pattern recognition. IEEE Computer Society Press, Washington, pp 864–868

Hsu C, Chang C, Lin C (2003) A practical guide to support vector classification. Department of Computer Science and Information Engineering, National Taiwan University. Available via http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pd cited 20 December 2006

Maenhout S, De Baets B, Haesaert G, Van Bockstaele E (2007) Marker-based screening of maize inbred lines using support vector machine regression. Euphytica doi: 10.1007/s10681-007-9423-5 (in press)

Melchinger A (1999) Genetic diversity and heterosis. In: Coors J, Pandey S (eds) The genetics and exploitation of heterosis in crops. American Society of Agronomy, Madison, pp 99–118

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

Smola A, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222

Stuber C, Cockerham C (1966) Gene effects and variances in hybrid populations. Genetics 54:1279–1286

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP–a new technique for DNA-fingerprinting. Nucleic Acids Res 23:4407–4414

Vuylsteke M, Mank R, Antonise R, Bastiaans RE, Senior M, Stuber C, Melchinger A, Lübberstedt T, Xia X, Stam P, Zabeau M, Kuiper M (1999) Two high-density AFLP (R) linkage maps of

Zea mays L.: analysis of distribution of AFLP markers. Theor Appl Genet 99:921–935

Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V (2000) Feature selection for SVMs. In: Advances in neural information processing systems. vol 13. MIT, Cambridge, pp 668–674

Wright S (1978) Variability within and among natural populations. In: Evolution and the genetics of populations. vol. 4, University of Chicago Press, Chicago, pp 449–450

Zien A, Ratsch G, Mika S, Scholkopf B, Lengauer T, Muller KR (2000) Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 16(9):799–807