Improved modeling of clinical data with kernel methods

Daemen, Anneleen; Timmerman, Dirk; Van den Bosch, Thierry; Bottomley, Cecilia; Kirk, Emma; Van Holsbeke, Caroline; Valentin, Lil; Bourne, Tom; De Moor, Bart

Published in:
Artificial Intelligence in Medicine

DOI:
10.1016/j.artmed.2011.11.001

2012

Improved modeling of clinical data with kernel methods

Anneleen Daemen[a], Dirk Timmerman[b], Thierry Van den Bosch[b], Cecilia Bottomley[c], Emma Kirk[d], Caroline Van Holsbeke[b,e], Lil Valentin[f], Tom Bourne[b,g], Bart De Moor[a]

[a]Department of Electrical Engineering, Katholieke Universiteit Leuven, 3001 Leuven, Belgium; [b]Department of Obstetrics and Gynecology, University Hospitals Leuven, Katholieke Universiteit Leuven, 3000 Leuven, Belgium; [c]Department of Obstetrics and Gynaecology, St George's Hospital, St George's University of London, London SW17 0RE, UK; [d]Early Pregnancy and Gynecological Unit, St George's Hospital, St George's University of London, London SW17 0RE, UK; [e]Hospital Oost-Limburg, 3600 Genk, Belgium; [f]Malmö University Hospital, Lund University, SE 20502 Malmö, Sweden; [g]Hammersmith Hospital, Imperial College London, London W12 0NN, UK.

Correspondence:
Anneleen Daemen, Ph.D.
Department of Cancer & DNA Damage Responses
Life Sciences Division
Lawrence Berkeley National Laboratory
One Cyclotron Road, 94720 Berkeley
California, USA
e-mail: anneleen.daemen@gmail.com
tel: 1 (510) 486 5202

Abstract

Objective

Despite the rise of high-throughput technologies, clinical data such as age, gender and medical history guide clinical management for most diseases and examinations. To improve clinical management, available patient information should be fully exploited. This requires appropriate modeling of relevant parameters.

Methods

When kernel methods are used, traditional kernel functions such as the linear kernel are often applied to the set of clinical parameters. These kernel functions, however, have their disadvantages due to the specific characteristics of clinical data, being a mix of variable types with each variable its own range. We propose a new kernel function specifically adapted to the characteristics of clinical data.

Results

The clinical kernel function provides a better representation of patients' similarity by equalizing the influence of all variables and taking into account the range *r* of the variables. Moreover, it is robust with respect to changes in *r*. Incorporated in a least squares support vector machine, the new kernel function results in significantly improved diagnosis, prognosis and prediction of therapy response. This is illustrated on four clinical data sets within gynecology, with an average increase in test area under the ROC curve (AUC) of 0.023, 0.021, 0.122 and 0.019, respectively. Moreover, when combining clinical parameters and expression data in three case studies on breast cancer, results improved overall with use of the new kernel function and when considering both data types in a weighted fashion, with a larger weight assigned to the clinical parameters. The increase in AUC with respect to a standard kernel function and/or unweighted data combination was maximum 0.127, 0.042 and 0.118 for the three case studies.

Conclusion

For clinical data consisting of variables of different type, the proposed kernel function– which takes into account the type and range of each variable–has shown to be a better alternative for linear and non-linear classification problems.

Key words: machine learning; support vector machine; kernel function; biostatistics; clinical data representation; clinical decision support system; gynecology; breast cancer

1. Introduction

During an examination, patient-specific information such as age, menopausal status and medical history is registered. Histopathological parameters such as tumor size, lymph node status and relapse rate, and ultrasound data such as endometrium thickness are often registered as well, with the set of clinical parameters characterizing a patient depending on the investigated disease. Such parameters or combinations thereof have been evaluated as prognostic indicators (for example, [1,2]). Because clinicians prefer interpretable decision support systems, clinical management for diagnosis and prognosis and decisions

concerning therapy response are for most of the diseases and examinations fully based on clinical and pathological indicators.
Besides clinical data, high-throughput technology–and especially microarray technology–has considerably advanced basic biological science and the entire field of cancer taxonomy, biomarker development and identification of prognostic and predictive markers [3-5]. In numerous studies, multiple high-throughput data sources were collected and simultaneously studied while omitting clinical parameters. High-throughput data, however, are in general much more difficult and expensive to collect while clinical parameters are routinely measured by clinicians. The latter have been used by clinicians for decades and should be included in the investigation, moreover because a critical study on the prediction of breast cancer outcome has suggested that clinical markers and profiles obtained from high-throughput technologies have similar power for prognosis [6].
Advanced mathematical models can aid clinical decision support. In many previous studies [7-10], the support vector machine (SVM) [11] was used for this purpose. Several disadvantages, however, occur when applying the SVM directly to clinical data, due to the heterogeneous nature of clinical data compared to high-throughput data sources. The influence of each variable on patients' similarity will be proportional to its range, thereby enlarging the influence of irrelevant continuous variables and diminishing the contribution of important discrete variables. As it has been shown that better results can be obtained by adapting the kernel function to the structure of the data and defining a kernel function per domain [12], a distinction is made between continuous variables, ordinal variables with an intrinsic ordering but often lacking equal distance between two consecutive categories, and nominal variables without any ordering.
The scale of the input data was already known to influence model performance. A rough distinction according to variable type was incorporated in LS-SVMlab, a Matlab/C toolbox containing a variety of techniques and algorithms for the least squares support vector machine (LS-SVM) with applications in classification and non-linear regression [13]. Binary variables were re-scaled to {-1,1} whilst continuous variables were normalized, avoiding attributes in larger numeric ranges to dominate those in smaller ranges. Other variables, however, were kept unchanged, thereby not distinguishing ordinal from nominal variables.
We will propose an alternative kernel function specifically developed for clinical data, which does not suffer from the ambiguity of data preprocessing by equally taking into account all variables. First, we will show the improvement obtained with this alternative kernel function when applied to four clinical data sets within gynecology. Secondly, the advantage of this kernel function will be illustrated for the combination of clinical and microarray data in three case studies on breast cancer.

2. Methods

2.1 Kernel methods and least squares support vector machine

Kernel methods are a powerful class of algorithms for pattern analysis. They work in a high dimensional feature space to which data x is mapped from the original input space with the function $\Phi(x)$ [14,15]. The kernel function $k(x^i, x^j)$ efficiently computes the inner product $\langle \Phi(x^i), \Phi(x^j) \rangle$ between all pairs of data items x^i and x^j in the feature space,

resulting in the *N x N* kernel matrix *K* with *N* the number of data items. Any symmetric, positive semi-definite function is a valid kernel function, resulting in many possible kernels. However, no formal proof of optimality exists for the use of one kernel function above an other. The functions that are most frequently employed in classification problems are the linear kernel $x^{i'} x^j$, the polynomial kernel $\left(x^{i'} x^j + \tau\right)^d$ with–as kernel parameters–the intercept constant $\tau \in \mathbb{R}^+$ and degree $d \in \mathbb{N}$, and the radial basis function $\exp\left(-\left\|x^i - x^j\right\|_2^2 / \sigma^2\right)$ with $\sigma \in \mathbb{R}^+$ representing the width of a Gaussian distribution centered on the data points. The polynomial kernel corresponds to a feature space spanned by all products of *d* variables at the most. This kernel results in a quadratic separating surface in the input space for $d = 2$, and it represents the cubic kernel for $d = 3$. More complex kernel functions have been proposed as well, such as graph and wavelet kernels [16,17]. In this paper, the linear kernel function is compared with a newly introduced kernel function for clinical data, referred to as the *clinical kernel function* (see section 2.3).

A kernel algorithm for supervised classification is the LS-SVM, a simplified version of the SVM [11] and developed by Suykens *et al.* [18,19]. Given is a training set for classification $\left\{x^i, y_i\right\}_{i=1}^N$ of *N* samples with feature vectors $x^j \in \mathbb{R}^p$ and binary output labels $y_i \in \{-1,+1\}$. The aim of supervised classification is to train a function $f(x) = y$ that correctly classifies unseen samples $\{x,y\}$. Data points x^i with $f\left(x^i\right) \geq 0$ are assigned the label +1, data points with $f\left(x^i\right) < 0$ the label –1. A non-linear function of the form $f(x) = w^T \Phi(x) + b$, with *w* representing the normal vector on the decision hyperplane $w^T \Phi(x) + b = 0$ and variable *b* the bias term, can be obtained with the following constrained optimization problem for the LS-SVM:

$\min_{w,b,e}\left(\frac{1}{2} w^T w + \gamma \frac{1}{2}\sum_{i=1}^N \zeta_i e_i^2\right)$ subject to $y_i\left[w^T \Phi\left(x^i\right) + b\right] = 1 - e_i \quad i = 1..N$ with

$\zeta_i = \begin{cases} \frac{N}{2N_P} & \text{if} \quad y_i = +1 \\ \frac{N}{2N_N} & \text{if} \quad y_i = -1 \end{cases}$, and N_P and N_N representing the number of positive and negative samples, respectively.

The regularization parameter γ represents the trade-off between maximization of the distance between samples of the two considered classes (that is, $2/\|w\|_2$) and minimization of the squared error contribution. Regularization by keeping γ small allows tackling the problem of overfitting by enforcing low complexity and good generalizability while tolerating misclassifications in case of overlapping distributions. Because in many two-class problems data sets are skewed in favor of one class with $N_P \gg N_N$ or $N_N \gg N_P$, we used an adapted version of the LS-SVM in which a different

factor ξ_i is assigned to positive and negative samples [20]. In this way, the contribution of false negative and false positive errors to the objective function is balanced.

In dual space, the equivalent problem of this optimization problem is a system of linear equations in function of the number of samples [18,19]. All experiments and calculations in this study were therefore performed in dual space, using Matlab 7.0.0 for Windows.

2.2 Kernel-based integration of multiple data sets

The representation of any data set with a real-valued kernel matrix, independent of the nature or complexity of the data to be analyzed, makes kernel methods ideally positioned for heterogeneous data integration. In [21], Daemen and colleagues investigated whether clinical and microarray data can be efficiently combined. In most microarray studies on cancer, the focus is on the microarray analysis while clinical data are not modeled in the same manner. When integrating both heterogeneous data sources, advantage can be taken from the strength of both data sources. This approach has been improved and extended towards the inclusion of multiple high-throughput data sources [22]. Three ways to simultaneously learn from multiple data sources were discussed, differing in the stage of the model building process at which integration occurs and referred to as early, intermediate and late integration [21]. With early integration, the microarray and clinical data sets would be concatenated before model building. Due to the huge amount of genes, clinical variables would need to be very significant before being selected. The late integration approach in which the two sets of variables would be treated separately before combining the resulting classifiers at the end may fail in improving performance with respect to the separate models due to the high correlation between microarray and clinical data. We therefore opted for intermediate integration in which the data sets are treated as separate entities and then combined at the kernel level–possibly weighted as $\mu K_{CL} + (1-\mu) K_{MA}$ –before building one final model.

2.3 Kernel function for clinical data

A normalized kernel function provides a measure of similarity between patients based on their clinical profiles. Obtained similarity values with the normalized linear and polynomial kernel function, however, strongly depend on the range of each variable, favoring continuous variables with a large range (for example, age from 20 to 50 years contrary to progesterone from 0 to 5 nmol/l). Also for ordinal variables the comparison of two patients with value 1 and 2 depends on the range of this variable. These patients will be less similar when the variable has only three categories compared to six. Furthermore, when an ordinal variable equals zero, the inner product will always be zero, independent of patients' dissimilarity. For nominal variables that lack an intrinsic ordering, the inner product between two patients should only be larger than zero in case both patients have the same category.

In this manuscript, we introduce an alternative kernel function, the *clinical kernel function*, specifically developed for clinical data. A distinction is made between continuous, ordinal and nominal variables, and per variable type a kernel function is defined. To guarantee the same influence of each variable, the appropriate kernel function is applied to each variable individually before calculating the global, heterogeneous kernel matrix.

The following notations are used: $k(z_i, z_j)$ denotes the kernel function for variable z between patients i and j; $K_z(i,j) \forall i,j$ represents the corresponding individual kernel matrix for variable z; and $K(i,j) \forall i,j$ represents the global, heterogeneous kernel matrix. For clinical studies in which the data are non-linear separable, the polynomial version of the clinical kernel function is used, obtained by replacing $x^{i'} x^j$ in the polynomial kernel function by the clinical kernel definition (in the tables and figures referred to as *clin poly*).

Continuous and ordinal clinical variables

The ordinal variables in the considered data sets (see section 3.1 and 3.2) are bleeding score, color score, tumor stage, tumor grade and nodal status. For those variables, the categories were replaced by their rank. Under the assumption of an equal distance between two consecutive categories, the same kernel function is proposed for continuous and ordinal variables:

$$k(z_i, z_j) = \frac{r - |z_i - z_j|}{r},$$

with constant r the range of a continuous variable z or the number of categories minus 1 for an ordinal variable z. The value for r can be extracted from the data or can be based on clinical knowledge or *a priori* information from specialist literature. The difference in z-value between two patients i and j is compared with and rescaled to this range. When r is based on the training data, the test data may contain more extreme values for certain variables. However, the kernel matrix will remain positive semi-definite with only negative values besides the diagonal, expressing more dissimilarity with the training cases.

Nominal clinical variables

For nominal variables, the kernel function between patients i and j is defined as the Kronecker delta function. This corresponds to setting the smoothing parameter λ of the Aitchison and Aitken kernel method for unordered categorical data [23] to 1, eliminating the problem of choosing a suitable value for λ:

$$k(z_i, z_j) = \begin{cases} 1 & \text{if} \quad z_i = z_j \\ 0 & \text{if} \quad z_i \neq z_j \end{cases}.$$

This kernel function is independent of the variable values, making binary dummy variables obsolete.

Final kernel for clinical data

For each individual kernel function k, the similarity measure is forced to the interval [0,1]. The global, heterogeneous kernel matrix K can therefore be defined as the sum of the individual kernel matrices K_z divided by the total number of clinical variables p. K describes the similarity for a group of patients based on a set of variables of different type. This corresponds to the additive kernel function $k(x^i, x^j) = \frac{1}{p} \sum_{z=1}^{p} k(z_i, z_j)$.

Example
To illustrate the clinical kernel function, we calculate the similarity (that is, kernel matrix) between three patients *h*, *i* and *j* for the continuous variable age. Patient *h* is 23 years old, patient *i* is 26 and patient *j* 54. Suppose based on the training data that the minimal age is 20 and maximal age 100. The elements in the kernel matrix can then be calculated as follows:

$$k(\text{age}_h, \text{age}_i) = (80 - |23 - 26|) / 80 = 77 / 80$$
$$k(\text{age}_h, \text{age}_j) = (80 - |23 - 54|) / 80 = 49 / 80$$
$$k(\text{age}_i, \text{age}_j) = (80 - |26 - 54|) / 80 = 52 / 80$$

The resulting kernel matrix for variable age equals

$$K_{age} = \begin{bmatrix} 1 & 0.96 & 0.61 \\ 0.96 & 1 & 0.65 \\ 0.61 & 0.65 & 1 \end{bmatrix}.$$

The extent of most types of cancer is described with a TNM classification system: T represents the size of the primary tumor, with suppose ranks 1, 2, 3 and 4 for illustrative purpose; N describes the degree of spread of the tumor to regional lymph nodes (0, 1 or 2); the absence or presence of metastasis is represented by the binary variable M. Patient *h* is characterized by T1N0M0, patient *i* by T3N2M1 and patient *j* by T4N1M1. The resulting individual matrices are

$$K_T = \begin{bmatrix} 1 & 0.33 & 0 \\ 0.33 & 1 & 0.66 \\ 0 & 0.66 & 1 \end{bmatrix}, K_N = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}, K_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

This example illustrates that the kernel values decrease with increasing dissimilarity between patients. The proposed kernel function takes into account the range of variables (for example, 0.66 (T) vs. 0.5 (N) for a difference of one unit in the number of categories). Moreover, the kernel value equals zero when two patients are most dissimilar (T1 vs. T4 and N0 vs. N2). The linear kernel function, on the other hand, would have led to erroneous positive values.
The global, heterogeneous kernel matrix for the similarity between patients *h*, *i* and *j* based on age, tumor size (T), lymph node spread (N) and metastasis (M) is given by

$$K = \frac{1}{4}\left(K_{age} + K_T + K_N + K_M\right) = \begin{bmatrix} 1 & 0.32 & 0.28 \\ 0.32 & 1 & 0.70 \\ 0.28 & 0.70 & 1 \end{bmatrix}.$$

A comparison of the proposed clinical kernel function with the linear and polynomial functions on real data sets is provided in section 4.

3. Experiments

3.1 Clinical data

We considered four clinical data sets. A binary outcome was selected or constructed for prediction, and a distinction was made between continuous variables (labeled as C), ordinal variables (O), and nominal variables (N). Before the analyses, some of the original variables were log-transformed. To avoid deteriorating performances, each data set was also reduced to a set of variables lacking redundancy by investigating the pairs of variables with a Spearman correlation coefficient in absolute value above 0.7. Details of the eligibility criteria, patient information registration and examinations have been published in the respective original publications.

I) Endometrial disease: abnormal versus normal

Data set I contains clinical information on 402 patients with an endometrial disease who underwent an ultrasound examination and color Doppler [24]. The patients were divided into two groups according to their histology: abnormal (hyperplasia, polyp, myoma, and carcinoma) versus normal (proliferative endometrium, secretory endometrium, atrophia). After excluding patients with incomplete data and correlation-based exclusion of redundant variables, data set I contained 22 variables for 339 patients, of which 163 were abnormal and 176 normal. An overview of the 22 clinical variables is given in Table 1.

II) Miscarriages: miscarriage versus vital fetus

A prospective observational study of 1828 women undergoing transvaginal ultrasound before 12 weeks gestation resulted in data for 2356 pregnancies. Among them, 1458 were normal at week 12 whereas 898 had miscarried by the end of week 12 [25]. The 18 clinical variables are shown in Table 2.

III) Pregnancies of unknown location (PUL): ectopic pregnancy (EP) versus other types of PUL

Data set III contains data on 1003 PULs [26]. Both persisting PULs (18 cases) and pregnancies with missing data were excluded, resulting into 856 PULs among which there were 460 failing PULs, 330 intrauterine pregnancies and 66 EPs. Because correct classification of EPs among PULs has been shown to be the most important diagnostic problem [27], the 66 EPs were considered versus the 790 other PULs. We refer to Table 3 for an overview of the 12 clinical variables.

IV) Adnexal masses: malignant versus benign

As fourth clinical data set, we studied a multicentric data set on adnexal masses, collected by the international ovarian tumor analysis group (IOTA) during phase 1 and 1b [2,28]. More than 40 clinical and ultrasound variables were collected from 1573 patients, of whom 1164 had a benign and 409 a malignant adnexal mass. We considered only those ultrasound variables that were included in at least one of the previously developed models for the calculation of the risk of malignancy in adnexal tumors (see Table 4) [28].

3.2 Clinical and expression data

Besides pure clinical data sets, we also considered three case studies in which both microarray data and a sufficient number of clinical variables were available. The microarray data were obtained with the Affymetrix technology and preprocessed with MAS5.0 (Affymetrix). An updated array annotation was used for the conversion of probes to entrez gene ids [29]. Finally, genes with low variation were excluded in an unsupervised way, retaining the 5000 genes with the largest standard deviation across all samples.

V) Breast cancer: recurrence
The first data set of 129 patients contained information on 17 clinical variables [30]. After exclusion of redundant variables, variables with too many missing values and patients with missing clinical information, this data set (referred to as data set V and represented in Table 5) consisted of 110 patients, in 85 of whom disease didn't recur whilst in 25 patients disease recurred.
VI) Breast cancer: treatment response
The second data set, in which response to treatment was studied, consisted of 12 variables for 133 patients [31]. Patient and variable exclusion as described above resulted in data set VI. Of the 129 remaining patients, 33 showed complete response to treatment while 96 patients were characterized as having residual disease. An overview of the 8 variables is provided in Table 6.
VII) Breast cancer: relapse
In the last case study, relapse was studied in 187 patients [32]. After preprocessing, data set VII retained information on 5 variables for 177 patients. In 112 patients, no relapse occurred while 65 patients were characterized as having a relapse. We refer to Table 7 for an overview of the variables included in this case study.

3.3 Model building strategy

For the clinical data sets (section 2.4), the data were randomly split into $2/3^{rd}$ for training and $1/3^{rd}$ for testing. This split was performed stratified to outcome to ensure that the relative proportion of outcomes sampled in both training and test set was similar to the original proportion in the full data set. On the training samples, a 10-fold cross-validation (CV) approach was applied for the optimization of the regularization parameter γ on a logarithmic scale from 10^{-4} to 10^{+6}. When the polynomial kernel function was used, a three-dimensional grid was required for the additional optimization of tuning parameters τ and d, both varying on a linear scale from 1 to 5. Contrary to the typical use of the polynomial kernel with $\tau = 1$ [33-35], scaling (that is, $\tau \neq 1$) was considered as this has shown to increase test performance [36]. The optimal parameter values were chosen corresponding to the model with the highest 10-fold train area under the receiver operating characteristic curve (AUC). In case of multiple models with equal AUC, the model with the lowest balanced error rate and an as high as possible sum of sensitivity and specificity was chosen. This optimal model was further validated on the $1/3^{rd}$ of samples left out for testing. To obtain a better estimate of the prediction performance of the classifiers and thus a more reliable comparison of the clinical alternative with the traditional kernel functions, the split of the data in a training and test part was repeated 100 times. The 10-fold train AUC and test AUC values averaged over these 100 repetitions are reported. The one-sided paired-sampled t-test was used to compare the AUC values obtained with the applied kernel functions. A p-value of 0.05 was considered statistically significant. For the linear and the polynomial kernel function, the more robust normalized version $k(x^i, x^j) = k(x^i, x^j) / \sqrt{k(x^i, x^i) k(x^j, x^j)}$ was used, although not explicitly mentioned in the remaining of the paper. To fairly compare the kernel functions, the linear and polynomial kernel functions were applied to both the raw data (in the tables and figures referred to as *linear/poly*) as well as the data after normalization (referred to as *linear norm/poly norm*). For the normalization, continuous and ordinal

variables were re-scaled to a range of 0 to 1, whilst nominal variables with k categories were replaced by $k-1$ binary dummy variables.

For the three clinical data sets for which corresponding expression data were available (section 2.5), the number of available samples was much smaller (in the order of hundreds as compared to thousands of samples in pure clinical studies). The intercept constant τ in the polynomial kernel function was therefore fixed to 1. Moreover, instead of selecting an independent test set, 10-fold CV was applied to each full data set and repeated 100 times. For each repetition, the random division of data into 10 folds was performed with stratification to outcome. For the kernel matrix obtained from the microarray data set in each case study, the 200 most differentially expressed genes selected from the training data with the Wilcoxon rank-sum test were considered. The performance of the clinical kernel with respect to the traditional kernel functions was independent of the specific number of incorporated genes.
Because the normalized linear and polynomial kernel functions were used and the kernel values obtained with the clinical alternative lay between 0 and 1 due to their construction, the weights assigned to the kernel matrices reflect the importance of each individual data set for the problem at hand. In the case studies with clinical data (CL) and microarray data (MA), three settings were therefore considered for the evaluation of the clinical kernel function: only the clinical data sets were considered for classification (1 CL + 0 MA), the influence of microarray and clinical data on prediction was set equal as has shown to be sufficient under some assumptions [21,37,38] (½ CL + ½ MA), and the weights assigned to both data sets were optimized because equal weights are not optimal when both data sets are of different relevance [39] (μ CL + (1-μ) MA). In the latter, we made μ vary from 0 to 1 in steps of 0.05, and the combination (γ,μ) or (γ,μ,d) that led to the largest 10-fold AUC was selected. For comparison, the results obtained with only microarray data (o CL + 1 MA) are reported as well.
The models based on only clinical data were also compared to a conventional prognostic index, the Nottingham prognostic index (NPI). The formula for the NPI equals 0.2 x tumor size (cm) + tumor grade (1-3) + lymph node stage (1-3) [1], and could only be applied to data set V.

4. Results

4.1 Comparison of the linear and clinical kernel function

In a first phase, we verified whether the clinical kernel function better represents true similarity between patients. For this purpose, a publicly available data set on breast cancer containing a mix of continuous, ordinal and nominal variables was used in which the appearance of distant subclinical metastases was predicted based on the primary tumor [40]. The data set of 148 patients contained 13 clinical parameters, represented in Table 8 [41]: 2 continuous parameters, age (20-60 years old) and tumor diameter (0-70 mm); 4 ordinal parameters, one with 16 categories and 3 with 3 categories each; and 7 nominal, binary parameters.
Four comparisons based on the patients' data shown in Table 8 were made to verify whether differences in kernel values correspond to true differences in patient data. Patients 193 and 265 differ greatly in tumor size, are most different for the ordinal

variables and are distinct concerning all nominal variables. The clinical kernel function assigns to them a kernel value of 0.152, contrary to 0.141 for the linear kernel function. Patients 153 and 193, on the other hand, are most dissimilar according to the linear kernel function with a kernel value of 0.009, contrary to a clinical kernel value of 0.390. Although age and tumor size are more different compared to patients 193 and 265, patients 153 and 193 have the same characteristics for two ordinal and two nominal variables. The clinical kernel function ranks these patients as more similar because the influence of all variables is equalized. In the linear kernel, the influence of the continuous variables age and tumor size dominates the influence of the non-continuous variables. We subsequently validated the separate influence of continuous and non-continuous variables on the calculation of the kernel matrix when keeping the other variables fixed. Patients 4 and 109, for example, are different according to two nominal variables, whilst patients 4 and 174 slightly differ in two ordinal variables (0 vs. 1 for O_1 with 16 categories; 1 vs. 2 for O_2 with 3 categories). Taking into account the range of the variables, patients 4 and 174 are more similar than patients 4 and 109. This difference in similarity is much clearer with the clinical kernel function (0.956 and 0.846, respectively) than with the linear kernel function (0.946 and 0.931, respectively). The final 4 patients in Table 8 only differ in age and tumor size. For both patients 26 and 199 and patients 9 and 251, this difference is 1 year in age and 1 mm in tumor size, with the latter pair being older with a slightly larger tumor. For the clinical kernel function, the similarities $k(26,199)$ and $k(9,251)$ are both equal to 0.997. These similarities, however, are slightly different according to the linear kernel function (0.9983 and 0.9984).
These comparisons of the linear and clinical kernel function show that differences in kernel values obtained with the linear kernel function do not optimally reflect true differences in patient data. This is caused by continuous variables dominating non-continuous ones and because the range of the variables is not taken into account. Patients are assigned to be similar when only ordinal and nominal variables differ, or dissimilar when differing more with respect to the continuous variables. In general, we can conclude from these comparisons that the clinical kernel provides a better representation of patients' similarity by equalizing the influence of each variable and taking into account the range of the variables.

4.2 Results for clinical data

We compared the linear and polynomial kernel function with the clinical alternatives on four data sets when used in a supervised classification algorithm. A 10-fold CV approach was applied for training an LS-SVM model, subsequently validated on a test set. The 10-fold train and test results averaged over 100 random repetitions are shown in Table 9 while the corresponding boxplots are provided in Figure 1. Overall, the LS-SVM models based on the clinical kernel definition significantly outperformed the models based on the linear and polynomial kernel (with and without normalization of the data). The test performance obtained with the polynomial kernel function after data normalization was slightly better in 1 case compared to the non-linear clinical kernel function. For the linear clinical kernel, the increase in 10-fold train AUC values ranged from -0.008 to 0.24, whereas the interval was [-0.08; 0.28] for the test AUC values. For the non-linear version, the intervals were [-0.005; 0.28] and [-0.15; 0.29], respectively.

When comparing the polynomial kernel with its clinical variant, the intercept and degree were–in the majority of repetitions–1 for the polynomial kernel, while varying from 1 to 5 and 3 to 5, respectively for the non-linear clinical kernel. When comparing the clinical alternatives, in three out of the four case studies the non-linear clinical kernel outperformed the linear version on the training data; however, it produced worse test results for all four data sets. The linear clinical kernel function has thus a better generalization performance. Moreover, the clinical kernel function not only outperformed the linear and polynomial kernel when using the LS-SVM classifier. It also performed well in combination with the regular SVM classifier [42].
Without the intention to exhaustively compare the SVM and LS-SVM with other classification methods, we applied three widely used classifiers to the clinical data, being Naive Bayes, K-nearest neighbor and decision trees. For Naive Bayes, the normal distribution was used to model continuous variables, whilst ordinal and nominal variables were modeled with a multivariate multinomial distribution. Prior probabilities for the classes were estimated from the relative frequencies of the classes in the training data. For the K-nearest neighbor algorithm, the number of nearest neighbors used for classification was set to 3. Finally for the decision tree, the minimal number of samples per node and tree leaf was set to 10 and 1, respectively, and a distinction was made between continuous/ordinal and nominal variables. For all three methods, training on $2/3^{rd}$ of the samples and testing on $1/3^{rd}$ of the samples was repeated 100 times, with use of the same splits as for the LS-SVM. The average test accuracies for the three methods when applied to the 4 clinical data sets and for the LS-SVM with use of the best clinical kernel function are shown in Table 10. The LS-SVM performed better for 2 out of 4 data sets and similar than at least one of the three other approaches for the other 2 data sets.

4.3 Results for clinical and expression data

The results for the three case studies with clinical and microarray data are shown in Table 11 and Figure 2. When only clinical data were considered, the same trend as with the 4 previous data sets was observed, that is, a significant increase in performance was obtained with the clinical kernel definition. Applying the NPI to data set V resulted in an AUC of 0.604, which was worse than both the traditional kernel functions (AUC = 0.782-0.793) and the clinical kernel functions (AUC = 0.818). When combining the clinical data with microarray data, the clinical kernel variant resulted in a significant improvement for the three data sets, both for an equal influence of clinical and microarray data and with the weights assigned to both data sets optimized (only the latter for data set V). Overall, the clinical variant of the polynomial kernel performed slightly better than the linear clinical kernel, likely due to the complexity of the classification problems caused by the heterogeneity of breast cancer and the low number of samples. For data set V and VII, however, a degree of 1 already led to optimal results, both for μ optimized and set to 0.5. Results are shown when the 200 most differential genes were considered for the calculation of the microarray-based kernel matrix. A similar trend was observed with the inclusion of less (20, 50, 100) and more (500) genes.
Tables 5, 6 and 7 contain for the clinical parameters the univariate results, which differ in function of the predicted outcome. Age, estrogen and progesterone status are more important for treatment response whilst tumor stage and size are important factors for the prediction of recurrence and relapse. Figure 3 shows per case study the histogram of the

weights assigned to the clinical data set when combined with microarray data for the linear and clinical kernel function. Similar distributions were obtained for the polynomial kernel function and its clinical variant. A clear link was observed between the weights assigned to the kernel matrix for the clinical data set when based on the linear kernel function before data normalization and the significance of the continuous variables (age, tumor size). The influence of continuous variables with a wide range is much larger on the calculated patients' similarities than that of ordinal and nominal variables. When these continuous variables are in addition significantly related to the predicted outcome, the corresponding kernel matrix is assigned a large weight. This was the case for data sets V and VII caused by the relevance of tumor size. For data set VI weights were spread between 0 and 1 due to the limited relevance of age, the only variable with a large range. After normalization of the data on the other hand, smaller weights were assigned to the data with use of the linear kernel function compared to the clinical kernel function.
We also investigated the effect of data integration on performance. Compared to the use of only clinical data (1 CL + 0 MA) or microarray data (0 CL + 1 MA; Table 11), a better performance was obtained for all kernel functions in all three case studies when considering both clinical and microarray data with the weights assigned to them optimized (μ CL + (1- μ) MA). When equal weights were assigned to both data sets (½ CL + ½ MA), the performance for the linear and polynomial kernel function decreased. Moreover, the histograms in Figure 3 show that in the majority of repetitions a larger weight was assigned to the clinical data than to the microarray data. Whether clinical data are sufficient and whether the weights should be optimized, however, is often not known beforehand and depends on the specific application and data sets. In [19], the data integration approach was applied to another breast cancer data set for which weight optimization was not beneficial, neither was microarray data with respect to the available clinical parameters.

4.4 Robustness of the clinical kernel function

The clinical kernel function depends on one parameter that needs to be set in advance: for continuous variables the range; for ordinal variables the number of categories. This parameter *r* can be based on the training data or on experience or literature information. We investigated the robustness of the clinical kernel function to changes in *r*. The distances $|z_i - z_j|$ for variable *z* between patients *i* and *j* vary from 0 to *r*. When the range for *z* is based on the training data, the kernel values vary from 0 to 1. When enlarging the range based on experience or literature information, the kernel values will vary between a positive number smaller than 1 and 1, thereby diminishing the richness of the kernel function, and possibly its ability to properly predict a patient's label. When decreasing the range, kernel values can become negative. However, using a value for *r* that deviates from the range based on the training data has only a small influence on performance. The clinical kernel function is therefore robust with respect to changes in the parameter *r*.
For the continuous variable age, we divided the case studies in three groups according to the application: the pregnancy-related case studies II and III; the uterus-related case studies I and IV; and the breast cancer case studies V, VI and VII. For case studies II and III, the difference in age range (15-48 vs. 14-49 years) had no influence on performance. For case study I, the original age range (22-85 years) was enlarged to the age range of case study IV (that is, 9-94 years). Also for case studies V, VI and VII, the previous age

ranges were replaced by their union (28-88 years). For case studies I, VI and VII, enlarging the range for age only caused a small decrease in performance (average AUC decrease = 0.0006), whilst for case study V, a small increase in AUC of 0.0001 was observed.

5. Discussion and conclusions

When applying the normalized linear or polynomial kernel function for modeling clinical data, the influence of each variable is proportional to its range. When mainly continuous variables with a large range are informative for the target outcome, good results are obtained with the linear kernel function. On the other hand when mainly ordinal and nominal variables with a small number of categories are relevant, the performance of the traditional kernel functions is poor as well. Correlation with outcome is often unknown beforehand, nominal variables with numerous categories can distort the calculation of patient similarity, and moreover, dependency on variable range should be discarded. Hence, each variable should have the same influence on the calculation of patient similarities, which was previously not the case. We therefore proposed a linear additive kernel function as alternative for the linear and polynomial kernel function that takes into account the type and range of each variable. This requires the specification of each type of variable, as well as the range for continuous variables and the number of categories for ordinal variables based on the training data or *a priori* knowledge. The clinical kernel definition is robust with respect to the specific choice of the range or number of categories.

From our results, we can conclude that the clinical kernel function represents similarities between patients more accurately. Moreover, the LS-SVM based on the clinical kernel variant significantly outperformed the linear and polynomial kernel function when tested on four pure clinical data sets and three sets of clinical parameters collected in microarray studies. When in the latter case studies expression data were added by using a kernel-based integration approach, the clinical kernel variant led to a significant increase in performance for the 3 case studies. Finally, the kernel function proposed in this paper is not limited to clinical data. Any data set consisting of different types of variables can benefit from this function. Moreover, the proposed kernel function can be used in combination with any kernel method or method that can be kernelized.

Acknowledgements

AD is research assistant of the Fund for Scientific Research - Flanders (FWO-Vlaanderen). BDM is full professor at the Katholieke Universiteit Leuven, Belgium. This work is partially supported by: 1. Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc & fellow grants. 2. Flemish Government: a. FWO: PhD/postdoc grants, projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0318.05 (subfunctionalization), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); b. IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-ovarian

tumors 070706 (IOTA3). 3. Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011). 4. EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] M. Galea, R. Blamey, C. Elston and I. Ellis. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat 22 (1992) 207-219.
[2] D. Timmerman, A.C. Testa, T. Bourne, E. Ferrazzi, L. Ameye, M. Konstantinovic, *et al*. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group. J Clin Oncol 23 (2005) 8794-8801.
[3] F.C. Geyer and J.S. Reis-Filho. Microarray-based gene expression profiling as a clinical tool for breast cancer management: are we there yet?. Int J Surg Path 17 (2009) 285–302.
[4] F. Cardoso, L. van't Veer, E. Rutgers, S. Loi, S. Mook and M.J. Piccart-Gebhart. Clinical application of the 70-gene profile: the MINDACT trial. J Clin Oncol 26 (2008) 729–735.
[5] J.A. Sparano. TAILORx: trial assigning individualized options for treatment (Rx). Clin Breast Cancer 7 (2006) 347–350.
[6] P. Edén, C. Ritz, C. Rose, M. Fernö and C. Peterson. "Good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. Eur J Can 40 (2004) 1837-1841.
[7] C. Lu, T. Van Gestel, J.A.K. Suykens, S. Van Huffel, I. Vergote and D. Timmerman. Preoperative prediction of malignancy of ovarium tumor using least squares support vector machines. Artif Intell Med 28 (2003) 281-306.
[8] M. Adjouadi, N. Zong and M. Ayala. Multidimensional pattern recognition and classification of white blood cells using support vector machines. Part Part Syst Charact 22 (2005) 107-118.
[9] S.K. Majumder, N. Ghosh and P.K. Gupta. Relevance vector machine for optical diagnosis of cancer. Lasers Surg Med 36 (2005) 323-333.
[10] B. Van Calster, D. Timmerman, C. Lu, J.A.K. Suykens, L. Valentin, C. Van Holsbeke, *et al*. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. Ultrasound Obstet Gynecol 29 (2007) 496-504.
[11] V. Vapnik. Statistical Learning Theory (Wiley, New York, 1998).
[12] T. Gärtner, J.W. Lloyd and P.A. Flach. Kernels and distances for structured data. Mach Learn 57 (2004) 205-232.
[13] K. Pelckmans, J.A.K. Suykens, T. Van Gestel, J. De Brabanter, L. Lukas, B. Hamers, *et al*. LS-SVMlab: a Matlab/C toolbox for Least Squares Support Vector Machines. Posted at http://www.esat.kuleuven.be/sista/lssvmlab/ (Accessed: 18 April 2011).
[14] J. Shawe-Taylor and N. Cristianini. Kernel methods for pattern analysis (Cambridge University Press, Cambridge, 2004).
[15] A. Daemen and B. De Moor. Development of a kernel function for clinical data. Proc of Conf of IEEE Engineering in Medicine and Biology Society (EMBC) (2009) 5913-5917.
[16] F. Fouss, K. Francoisse, L. Yen, A. Pirotte and M. Saerens. An experimental investigation of graph kernels on a collaborative recommendation task. Proc of Int Conf on Data Mining (ICDM) (2006) 863-868.

[17] Q. Wu and R. Law. Complex system fault diagnosis based on a fuzzy robust wavelet support vector classifier and an adaptive Gaussian particle swarm optimization. Information Sciences 180 (2010) 4514-4528.
[18] J.A.K. Suykens and J. Vandewalle. Least Squares Support Vector Machine classifiers. Neural Processing Letters 9 (1999) 293-300.
[19] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle. Least Squares Support Vector Machines (World Scientific, Singapore, 2002).
[20] G.C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. Proc. of IJCNN (2006) 1661-1668.
[21] A. Daemen, O. Gevaert and B. De Moor. Integration of clinical and microarray data with kernel methods. Proc of Conf of IEEE Engineering in Medicine and Biology Society (EMBC) (2007) 5411-5415.
[22] A. Daemen, O. Gevaert, F. Ojeda, A. Debucquoy, J.A.K. Suykens, C. Sempoux, *et al*. A kernel-based integration of genome-wide data for clinical decision support. Genome Medicine 1 (2009) 39.
[23] J. Aitchison and C.G.G. Aitken. Multivariate binary discrimination by the kernel method. Biometrika 63 (1976) 413-420.
[24] T. Van den Bosch, A. Daemen, O. Gevaert and D. Timmerman. Mathematical decision trees versus clinician based algorithms in the diagnosis of endometrial disease. Ultrasound Obstet Gynecol 30 (2007) 412.
[25] C. Bottomley, A. Daemen, F. Mukri, A.T. Papageorghiou, E. Kirk, A. Pexsters, *et al*. Functional linear discriminant analysis: a new longitudinal approach to the assessment of embryonic growth. Hum Reprod 24 (2009) 278-283.
[26] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, *et al*. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. Hum Reprod 21 (2006) 1824-1831.
[27] G. Condous, E. Okaro, A. Khalid, D. Timmerman, C. Lu, Y. Zhou, *et al*. The use of a new logistic regression model for predicting the outcome of pregnancies of unknown location. Hum Reprod 19 (2004) 1900-1910.
[28] C. Van Holsbeke, B. Van Calster, A.C. Testa, E. Domali, C. Lu, S. Van Huffel, *et al*. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. Clin Cancer Res 15 (2009) 684-691.
[29] M. Dai, P. Wang, A.D. Boyd, G. Kostov, B. Athey, E.G. Jones, *et al*. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 33 (2005) e175.
[30] K. Chin, S. DeVries, J. Fridlyand, P.T. Spellman, R. Roydasgupta, W.L. Kuo, *et al*. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 10 (2006) 529-541.
[31] K.R. Hess, K. Anderson, W.F. Symmans, V. Valero, N. Ibrahim, J.A. Mejia, *et al*. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol 24 (2006) 4236-4244.

[32] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, *et al*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98 (2006) 262-272.
[33] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 97 (2000) 262–267.
[34] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16 (2000) 906–914.
[35] M. Xiong, X. Fang and J. Zhao. Biomarker identification by feature wrappers. Genome Res 11 (2001) 1878–1887.
[36] T. Van Gestel, J.A.K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, *et al*. Benchmarking least squares support vector machine classifiers. Mach Learn 54 (2004) 5–32.
[37] H. Wainer. Estimating coefficients in linear models: it don't make no nevermind. Psychological Bulletin 83 (1976) 213–217.
[38] T. De Bie, L.C. Tranchevent, L.M.M. van Oeffelen and Y. Moreau. Kernel-based data fusion for gene prioritization. Bioinformatics 23 (2007) i125–i132.
[39] G.R.G. Lanckriet, T. De Bie, N. Cristianini, M.I. Jordan and W.S. Noble. A statistical framework for genomic data fusion. Bioinformatics 20 (2004) 2626–2635.
[40] M. van de Vijver, Y. He, L. van't Veer, H. Dai, A. Hart, D. Voskuil, *et al*. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347 (2002) 1999-2009.
[41] H. Chang, D. Nuyten, J. Sneddon, T. Hastie, R. Tibshirani, T. Sorlie, *et al*. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 102 (2005) 3738–3743.
[42] S. Yu, T. Falck, A. Daemen, L.C. Tranchevent, J.A.K. Suykens, B. De Moor, *et al*. L2 norm multiple kernel learning and its applications to biomedical data fusion. BMC Bioinformatics 11 (2010) 309.

Tables

Table 1: Clinical variables data set I (endometrial disease)

variable	type	range	p-value
1. age (years)	C	22 – 85	<0.0001*
2. weight (kg)	C	45 – 160	0.026*
3. number of miscarriages/abortions	O	0 – 5	0.825*
4. parity	O	0 – 7	0.316*
5. menopausal status	N	1,2,3	0.02º
6. hormonal therapy	N	0,1,2,3,4	0.048º
7. intrauterine device	N	0,1,2	0.107º
8. type of AUB$^{\rho}$	N	1,2,3	0.946º
9. amount of AUB$^{\rho}$	N	1,2,3	0.323º
10. duration of AUB$^{\rho}$ (months)	C	0.41 – 4.80$^{\Psi}$	0.393*
11. endometrial cells	N	1,2,3	0.174º
12. endometrium thickness on US$^{\varepsilon}$ (mm)	C	0 – 3.91$^{\Psi}$	<0.0001*
13. intracavity fluid (mm)	C	0 – 8.7	0.503*
14. 3-layer pattern	N	1,2	0.482º
15. intracavity lesion	N	1,2,3	<0.0001º
16. subendometrial cyst	N	1,2	0.488º
17. endometrial cyst	N	1,2	0.031º
18. number of calcifications	O	0 – 10	0.012*
19. number of myoma	O	0 – 4	<0.0001*
20. ovary aspect	N	1,2	0.365º
21. presence of follicles	N	0,1	0.887º
22. pedicle sign	N	1,2,3	<0.0001º

$^{\rho}$ AUB, abnormal uterine bleeding

$^{\varepsilon}$ US, ultrasound

$^{\Psi}$ on a logarithmic scale after correction for a positively skewed distribution (*log(x+1)*)

* Wilcoxon rank-sum test

º Fisher's exact test

Table 2: Clinical variables data set II (miscarriages)

variable	type	range	p-value
1. age (years)	C	15 – 48	<0.0001*
2. PBAC bleeding score	O	0 – 4	<0.0001*
3. follow-up consent	N	0,1,2	0.019º
4. ethnicity	N	0,1,2,3,4,5,6	
5. regular dates	N	0,1,2	<0.0001º
6. gravida	O	1 – 12	0.714*
7. number of deliveries after 24 weeks	O	0 – 10	0.447*
8. number of terminated pregnancies	O	0 – 4	0.001*
9. number of early miscarriages	O	0 – 10	0.848*
10. number of PULs$^{\rho}$	O	0 – 1	0.174*
11. number of late miscarriages	O	0 – 5	0.461*
12. number of ectopic pregnancies	O	0 – 1	0.391*
13. previous chromosomal abnormalities	N	0,1	0.047º
14. bleeding$^{\Psi}$	N	0,1	0.0006º
15. pain$^{\Psi}$	N	0,1	<0.0001º
16. previous ectopic pregnancy$^{\Psi}$	N	0,1	0.183º
17. previous miscarriage$^{\Psi}$	N	0,1	<0.0001º
18. anxiety$^{\Psi}$	N	0,1	<0.0001º

$^{\rho}$ PUL, pregnancy of unknown location
$^{\Psi}$ indication for scan
* Wilcoxon rank-sum test
º Fisher's exact test

Table 3: Clinical variables data set III (pregnancies of unknown location)

variable	type	range	p-value
1. hCG$^{\rho}$ at 48h (U/l)	C	0 – 9.52$^{\psi}$	<0.0001*
2. progesterone at 48h (nmol/l)	C	0 – 5.52$^{\psi}$	0.008*
3. endometrium thickness (mm)	C	0.92 – 3.58$^{\varepsilon}$	0.651*
4. character of midline echo	N	0,1	0.58^{o}
5. free fluid in pouch of Douglas	N	0,1	0.737^{o}
6. gestational age (days)	C	2.30 – 4.61$^{\psi}$	0.212*
7. lower abdominal pain	N	0,1	0.157^{o}
8. vaginal bleeding	N	0,1,2	0.034^{o}
9. previous miscarriage	N	0,1	1^{o}
10. previous ectopic pregnancy	N	0,1	0.007^{o}
11. anxiety	N	0,1	0.715^{o}
12. age (years)	C	14 – 49	0.378*

$^{\rho}$ hCG, human chorionic gonadotropin

$^{\psi}$ on a logarithmic scale after correction for a positively skewed distribution (*log(x)*)

$^{\varepsilon}$ on a logarithmic scale after correction for a positively skewed distribution (*log(x+1)*)

* Wilcoxon rank-sum test

^{o} Fisher's exact test

Table 4: Clinical variables data set IV (adnexal masses)

variable	type	range	p-value
1. age (years)	C	9 – 94	<0.0001*
2. personal history of ovarian cancer	N	0,1	0.004º
3. hormonal therapy	N	0,1	0.006º
4. maximal diameter of the lesion (mm)	C	2.08 – 6.02$^{\Psi}$	<0.0001*
5. presumed ovarian origin of tumor	N	0,1	0.16º
6. pelvic pain during examination	N	0,1	0.003º
7. locularity of the tumor (morphology of the lesion)	N	1,2,3,4,5,6	
8. maximal diameter of the solid component (mm)	C	0 – 50$^{\rho}$	<0.0001*
9. number of papillary projections	O	0 – 4	<0.0001*
10. blood flow within papillary projection	N	0,1	<0.0001º
11. irregular internal cyst walls	N	0,1	<0.0001º
12. acoustic shadows	N	0,1	<0.0001º
13. color score of intratumoral blood flow	O	1 – 4	<0.0001*
14. presence of venous blood flow only	N	0,1	0.004º
15. presence of ascites	N	0,1	<0.0001º

$^{\Psi}$ on a logarithmic scale after correction for a positively skewed distribution (*log(x)*)
$^{\rho}$ maximal diameter of the solid component bounded to 50mm due to its binomial distribution (that is, the diameter equals 0 in those patients without a solid component)
* Wilcoxon rank-sum test
º Fisher's exact test

Table 5: Clinical variables data set V (breast cancer – recurrence)

variable	type	range	p-value
1. age (years)	C	31 – 88	0.547*
2. ethnicity	N	0,1,2	0.151º
3. ER$^{\rho}$ status	N	0,1	0.049º
4. PR$^{\varepsilon}$ status	N	0,1	0.2º
5. radiation treatment	N	0,1	0.093º
6. chemotherapy	N	0,1	0.533º
7. hormonal therapy	N	0,1	0.674º
8. nodal status (N)	O	0 – 2	0.0001*
9. metastasis (M)	N	0,1	0.0004º
10. tumor stage	O	1 – 4	0.0002*
11. tumor size (cm)	C	0.262 – 2.14$^{\psi}$	<0.0001*
12. tumor grade	O	1 – 3	0.06*

$^{\rho}$ ER, estrogen receptor

$^{\varepsilon}$ PR, progesterone receptor

$^{\psi}$ on a logarithmic scale after correction for a positively skewed distribution (*log(x+1)*)

* Wilcoxon rank-sum test

º Fisher's exact test

Table 6: Clinical variables data set VI (breast cancer – treatment response)

variable	type	range	p-value
1. age (years)	C	28 – 79	0.045*
2. ethnicity	N	0,1,2,3,4	0.579º
3. pretreatment tumor stage	O	1 – 4	0.742*
4. nodal status (N)	O	0 – 3	0.763*
5. nuclear grade	O	1 – 3	0.0005*
6. ER$^{\rho}$ status	N	0,1	<0.0001º
7. PR$^{\varepsilon}$ status	N	0,1	0.0006º
8. HER2$^{\Psi}$ status	N	0,1	0.037º

$^{\rho}$ ER, estrogen receptor
$^{\varepsilon}$ PR, progesterone receptor
$^{\Psi}$ HER2, human epidermal growth factor receptor 2
* Wilcoxon rank-sum test
º Fisher's exact test

Table 7: Clinical variables data set VII (breast cancer – relapse)

variable	type	range	p-value
1. age (years)	C	32 – 86	0.331*
2. tumor size (cm)	C	0 – 2.22[ψ]	0.0015*
3. nodal status	N	0,1	0.301º
4. ER[ρ] status	N	0,1	0.031º
5. tamoxifen treatment	N	0,1	0.197º

[ρ] ER, estrogen receptor

[ψ] on a logarithmic scale after correction for a positively skewed distribution (*log(x+1)*)

* Wilcoxon rank-sum test

º Fisher's exact test

Table 8: Specific patient data from [36]

	Patient ID	193	265	153	193	4	109	174	26	199	9	251
	Disease status [ψ]	0	0	1	0	0	1	0	0	0	0	0
C_1	age (years)	50	41	37	50	41	41	41	40	39	48	49
C_2	tumor size (mm)	8	45	50	8	20	20	20	14	15	15	16
O_1	nb pos lymph nodes	0	4	0	0	0	0	1	0	0	0	0
O_2	N (pN0, 1-3, ≥4)	1	3	1	1	1	1	2	1	1	1	1
O_3	grade	3	1	1	3	1	1	1	1	1	1	1
O_4	NIH[ρ] risk	1	3	3	1	3	3	3	3	3	3	3
N_1	mastectomy	0	1	1	0	0	1	0	0	0	0	0
N_2	estrogen receptor	1	0	0	1	1	0	1	1	1	1	1
N_3	chemotherapy	0	1	0	0	0	0	0	0	0	0	0
N_4	hormonal therapy	0	1	0	0	0	0	0	0	0	0	0
N_5	St. Gallen criterion	0	1	1	0	1	1	1	1	1	1	1
N_6	NIH[ρ] consensus	0	1	1	0	1	1	1	1	1	1	1
N_7	T (≤ 2 cm or > 2 cm)	0	1	1	0	0	0	0	0	0	0	0

[ρ] NIH, National Institutes of Health

[ψ] appearance of distant subclinical metastases based on the primary breast tumor: yes = 1, no = 0

Table 9: Results for 4 clinical data sets within gynecology

data set	kernel function	10-fold AUC (std)	p-value[o]	test AUC (std)	p-value[o]
I	linear	0.742 (0.023)	1.1e-46	0.750 (0.037)	3.3e-24
	linear norm*	0.770 (0.022)	1.7e-34	0.781 (0.038)	8.3e-7
	clinical	0.786 (0.023)		0.791 (0.038)	
	poly	0.731 (0.025)	2.3e-47	0.738 (0.039)	2.8e-20
	poly norm*	0.770 (0.022)	6.0e-21	0.781 (0.037)	0.665
	clin poly	0.780 (0.023)		0.780 (0.043)	
II	linear	0.752 (0.008)	3.8e-66	0.754 (0.014)	1.1e-43
	linear norm*	0.763 (0.008)	2.2e-48	0.763 (0.013)	2.3e-30
	clinical	0.777 (0.008)		0.778 (0.013)	
	poly	0.735 (0.009)	2.6e-99	0.737 (0.014)	4.7e-55
	poly norm*	0.762 (0.008)	1.5e-86	0.763 (0.014)	3.5e-21
	clin poly	0.820 (0.008)		0.773 (0.015)	
III	linear	0.677 (0.028)	1.6e-73	0.688 (0.052)	7.3e-43
	linear norm*	0.656 (0.031)	1.1e-75	0.661 (0.056)	2.1e-46
	clinical	0.819 (0.022)		0.815 (0.038)	
	poly	0.648 (0.030)	1.9e-60	0.658 (0.058)	3.6e-24
	poly norm*	0.663 (0.023)	3.8e-79	0.641 (0.051)	1.3e-26
	clin poly	0.834 (0.022)		0.754 (0.071)	
IV	linear	0.912 (0.006)	2.4e-98	0.911 (0.012)	1.5e-72
	linear norm*	0.937 (0.005)	3.5e-84	0.935 (0.010)	1.8e-55
	clinical	0.945 (0.004)		0.944 (0.009)	
	poly	0.904 (0.006)	1.2e-100	0.904 (0.012)	8.8e-67
	poly norm*	0.937 (0.005)	1.6e-82	0.935 (0.010)	0.0143
	clin poly	0.948 (0.004)		0.936 (0.011)	

[o] one-sided paired-sampled t-test for the comparison of the linear and polynomial kernel with the clinical alternative

* continuous and ordinal variables were re-scaled to a range of 0 to 1, and nominal variables with *k* categories were replaced by *k-1* binary dummy variables before applying the kernel function

Table 10: Classifier comparison on 4 clinical data sets within gynecology, with the best performing classifier(s) for each data set underlined

data set	**test accuracy (std) for Naive Bayes**	**test accuracy (std) for K-nearest neighbor**	**test accuracy (std) for decision trees**	**test accuracy (std) for LS-SVM with optimal clinical kernel***
I	71.6 (3.9)	60.4 (3.9)	64.3 (4.5)	72.5 (3.9) – clinical
II	73.5 (1.2)	75.0 (1.2)	69.0 (1.9)	74.8 (1.3) – clinical
III	91.9 (0.8)	91.6 (0.9)	88.8 (1.8)	92.2 (0.6) – clin poly
IV	87.3 (1.2)	84.5 (1.1)	84.9 (1.6)	90.5 (1.6) – clinical

* Accuracy obtained with cut-off = 0 on the LS-SVM outcome

Table 11: Results of data integration for 3 case studies on breast cancer

data set	kernel function	*1 CL + 0 MA*		*½ CL + ½ MA*		*μ CL + (1-μ) MA*		*0 CL + 1 MA*
		10-fold AUC (std)	**p-value[o]**	**10-fold AUC (std)**	**p-value[o]**	**10-fold AUC (std)**	**p-value[o]**	**10-fold AUC (std)**
V	linear	0.793 (0.015)	9.2e-34	0.724 (0.025)	1.1e-83	0.794 (0.013)	2.6e-57	0.729 (0.027)
	linear norm*	0.793 (0.015)	2.3e-44	0.835 (0.018)	0.998	0.840 (0.019)	8.7e-17	
	clinical	0.818 (0.015)		0.832 (0.021)		0.851 (0.017)		
	poly	0.782 (0.014)	1.6e-52	0.735 (0.024)	3.9e-77	0.793 (0.017)	4.3e-67	
	poly norm*	0.783 (0.016)	7.2e-52	0.841 (0.019)	0.697	0.841 (0.019)	1.1e-11	
	clin poly	0.818 (0.015)		0.840 (0.018)		0.851 (0.017)		
VI	linear	0.799 (0.009)	9.9e-39	0.813 (0.007)	1.1e-47	0.813 (0.007)	5.5e-41	0.815 (0.006)
	linear norm*	0.791 (0.010)	9.8e-65	0.809 (0.010)	4.7e-50	0.818 (0.006)	1.4e-35	
	clinical	0.813 (0.008)		0.828 (0.008)		0.829 (0.008)		
	poly	0.812 (0.012)	1.1e-10	0.813 (0.007)	7.2e-48	0.818 (0.009)	4.7e-32	
	poly norm*	0.792 (0.010)	1.5e-62	0.813 (0.010)	1.2e-41	0.820 (0.007)	1.8e-33	
	clin poly	0.819 (0.008)		0.832 (0.007)		0.833 (0.007)		
VII	linear	0.650 (0.010)	3.0e-18	0.643 (0.015)	1.0e-60	0.665 (0.017)	1.7e-66	0.642 (0.022)
	linear norm*	0.624 (0.010)	6.0e-53	0.694 (0.020)	0.021	0.694 (0.020)	4.0e-38	
	clinical	0.662 (0.012)		0.697 (0.020)		0.742 (0.023)		
	poly	0.657 (0.010)	9.1e-27	0.649 (0.017)	2.0e-63	0.685 (0.017)	4.6e-54	
	poly norm*	0.647 (0.013)	8.9e-33	0.674 (0.016)	1.4e-34	0.690 (0.017)	2.0e-45	
	clin poly	0.676 (0.014)		0.718 (0.022)		0.742 (0.023)		

[o] one-sided paired-sampled t-test for the comparison of the linear and polynomial kernel with the clinical alternative

* continuous and ordinal variables were re-scaled to a range of 0 to 1, and nominal variables with *k* categories were replaced by *k-1* binary dummy variables before applying the kernel function

Figure captions

Figure 1
Boxplots of the 10-fold train and the test AUC values obtained in 100 repetitions for 4 clinical data sets on (A) endometrial disease, (B) miscarriages, (C) pregnancies of unknown location, and (D) adnexal masses.

Figure 2
Boxplots of the 10-fold AUC values obtained in 100 repetitions for 3 case studies on breast cancer, for the prediction of (A) recurrence, (B) treatment response, and (C) relapse. For each case study, results are shown when only clinical data are considered (1 CL + 0 MA; top), when clinical and microarray data have the same influence on prediction (½ CL + ½ MA; middle), and when the weights assigned to both data sets are optimized (μ CL + (1-μ) MA; bottom).

Figure 3
Histogram of the weights μ assigned to the clinical data in 100 repetitions when combined with microarray data according to μ CL + (1-μ) MA for the prediction of (A) recurrence, (B) treatment response, and (C) relapse. Weights are shown when assigned to the kernel matrix obtained with the linear kernel function (blue), with the linear kernel function after data normalization (green), and with the clinical kernel function (brown).

Figure 1

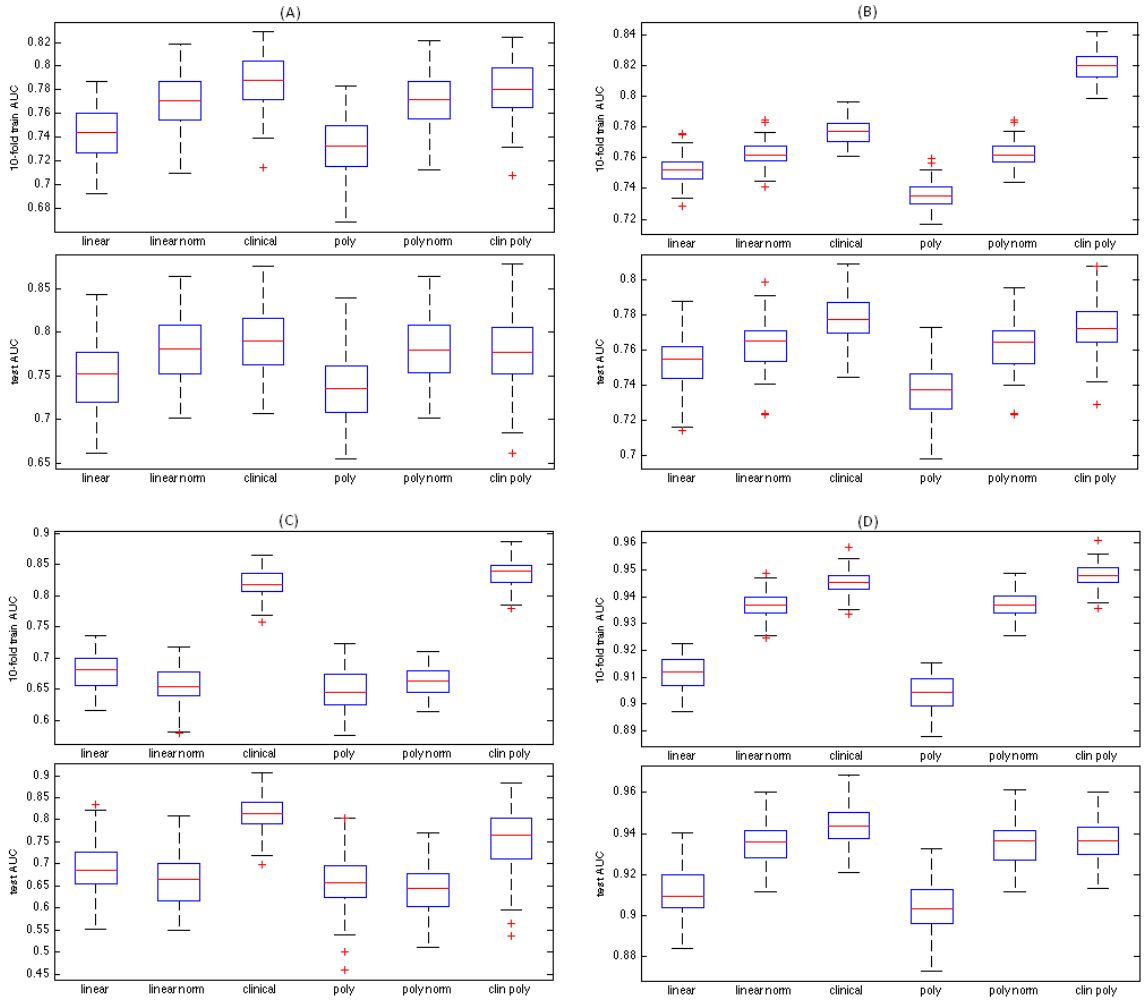

(A)
(B)
(C)
(D)
10-fold train AUC
test AUC
linear
linear norm
clinical
poly
poly norm
clin poly

Figure 2

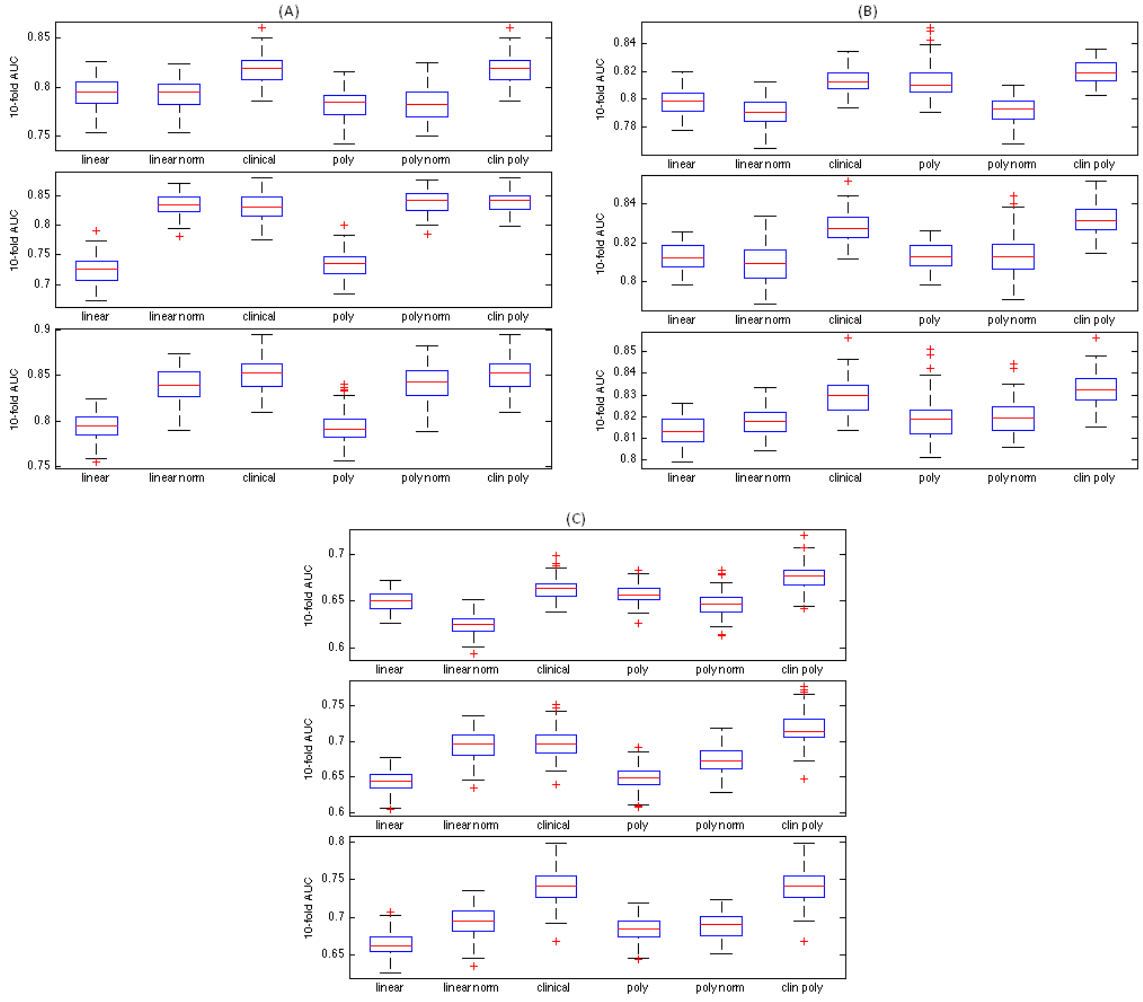
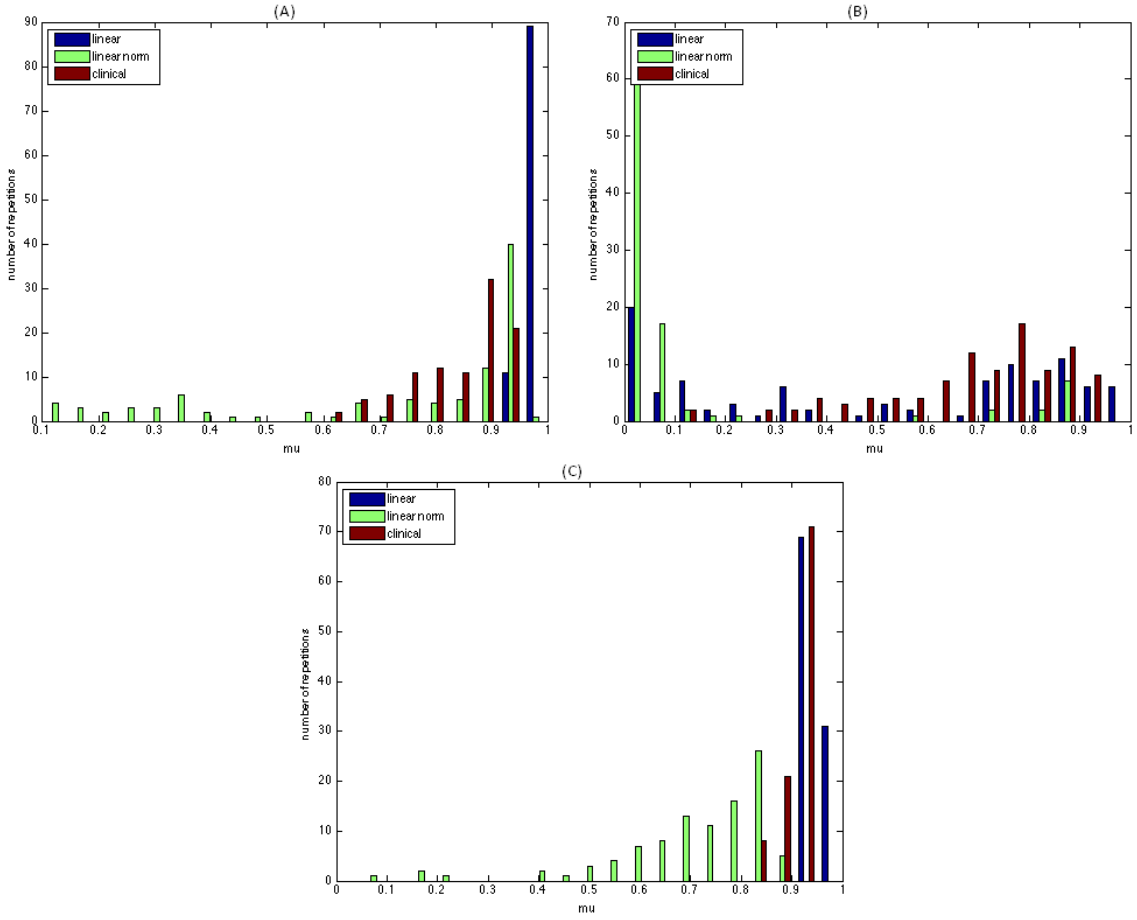

(A)
(B)
(C)
10-fold AUC
linear
linear norm
clinical
poly
poly norm
clin poly

Figure 3


(A)
linear
linear norm
clinical
number of repetitions
mu
(B)
linear
linear norm
clinical
number of repetitions
mu
(C)
linear
linear norm
clinical
number of repetitions
mu