Predicting breast cancer using an expression values weighted clinical classifier

Minta Thomas *[1], Kris De Brabanter[2], Johan AK Suykens[1]and Bart De Moor [1]

[1]KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics / iMinds Future Health Department, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

[2]Iowa State University, Department of Statistics Computer Science, Ames, IA

Email: minta.thomas@esat.kuleuven.be;

*Corresponding author

Abstract

Background: Clinical data, such as patient history, laboratory analysis, ultrasound parameters-which are the basis of day-to-day clinical decision support-are often used to guide the clinical management of cancer in the presence of microarray data. Several data fusion techniques are available to integrate genomics or proteomics data, but only a few studies have created a single prediction model using both gene expression and clinical data. These studies often remain inconclusive regarding an obtained improvement in prediction performance. To improve clinical management, these data should be fully exploited. This requires efficient algorithms to integrate these data sets and design a final classifier.

LS-SVM classifiers and generalized eigenvalue/singular value decompositions are successfully used in many bioinformatics applications for prediction tasks. While bringing up the benefits of these two techniques, we propose a machine learning approach, a weighted LS-SVM classifier to integrate two data sources: microarray and clinical parameters.

Results: We compared and evaluated the proposed methods on five breast cancer case studies. Compared to LS-SVM classifier on individual data sets, generalized eigenvalue decomposition (GEVD) and kernel GEVD, the proposed weighted LS-SVM classifier offers good prediction performance, in terms of test area under ROC Curve (AUC), on all breast cancer case studies.

Conclusions: Thus a clinical classifier weighted with microarray data set results in significantly improved diagnosis, prognosis and prediction responses to therapy. The proposed model has been shown as a promising mathematical framework in both data fusion and non-linear classification problems.

1 Background

Microarray technology, which can handle thousands of genes of several hundreds of patients at a time, makes it hard for scientists to manually extract relevant information about genes and diseases, especially cancer. Moreover this technique suffers from a low signal-to-noise ratio. Despite the rise of high-throughput technologies, clinical data such as age, gender and medical history, guide clinical management for most diseases and examinations. A recent study [1] shows the importance of the integration of microarray and clinical data has a synergetic effect on predicting breast cancer outcome. Gevaert et al. [2] have used a Bayesian framework to combine expression and clinical data. They found that decision integration, and partial integration leads to a better performance, whereas full data integration showed no improvement. These results were obtained by using a cross validation approach on the 78 samples in the van't Veer et al. [3] data set. On the same data set, Boulesteix et al. [4] employed random forests and partial least squares approaches to combine expression and clinical data. In contrast, they reported that microarray data do not noticeably improve the prediction accuracy yielded by clinical parameters alone.

The representation of any data set with a real-valued kernel matrix, independent of the nature or complexity of data to be analyzed, makes kernel methods ideally positioned for heterogeneous data integrations. Integration of data using kernel fusion is featured by several advantages. Biological data has diverse structures, for example, high dimensional expression data, the sequence data, the annotation data, the text mining data and heterogeneous nature of clinical data and so on. The main advantage is that the data heterogeneity is rescued by the use of kernel trick, where data which has diverse data structures are all transformed into kernel matrices with same size. To integrate them, one could follow the classical additive expansion strategy of machine learning to combine them linearly. These nonlinear integration methods of kernels have attracted great interests in recent machine learning research.

Daemen et al. [5] proposed kernel functions for clinical parameters and pursued an integration approach based on combining kernels (kernel inner product matrices derived from the separate data types) for application in a Least Squares Support Vector Machine (LS-SVM). They explained that the newly proposed kernel functions for clinical parameter does not suffer from the ambiguity of data preprocessing by equally considering all variables. That means, 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. They concluded that the clinical kernel functions represent similarities between patients more accurately than linear or polynomial kernel function for modeling clinical data. Pittman et al. [6] combined clinical and expression data for predicting breast cancer outcome by means of a tree classifier. This tree classifier was trained using meta-genes and/or clinical data as inputs. They explained that key metagenes can up to to a degree, replace traditional

risk factors in terms of individual association with recurrences. But the combination of metagenes and clinical factors currently defines models most relevant in terms of statistical fit and also, more practically, in terms of cross-validation predictive accuracy. The resulting tree models provide an integrated clinico-genomic analysis that generate substantially accurate and cross-validated predictions at the individual patient level.

Singular Value Decomposition (SVD) and generalized SVD (GSVD) have been shown to have great potential within bioinformatics for extracting common information from data sets such as genomics and proteomics data [7,8]. Several studies have used LS-SVM as a prediction tool, especially in microarray analysis [9, 10].

In this paper, we propose a machine learning approach for data integration: a weighted LS-SVM classifier. Initially we will explain generalized eigenvalue decomposition (GEVD) and kernel GEVD. Later we will explore the relationships of kernel GEVD with weighted LS-SVM classifier. Finally, the advantages of this new classifier will be demonstrated on five breast cancer case studies, for which expression data and an extensive collection of clinical data are publicly available.

2 Data sets

Breast cancer is one of the most extensively studied cancer types for which many microarray data sets are publicly available. Among them, we selected five cases for which a sufficient number of clinical parameters were available [3, 11–14]. All the data sets that we have used are available in the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA). Overview of all the data sets are given in Table 1.

2.1 Microarray Data

For the first three data sets, the microarray data were obtained with the Affymetrix technology and preprocessed with MAS5.0, the GeneChip Microarray Analysis Suite 5.0 software (Affymetrix). However, as probe selection for the Affymetrix gene chips relied on earlier genome and transcriptome annotation that are significantly different from current knowledge, an updated array annotation was used for the conversion of probes to Entrez Gene IDs, lowering the number of false positives [15].

A fourth data set consists of two groups of patients [3]. The first group of patients, the training set, consists of 78 patients of which 34 patients belonged to the poor prognosis group and 44 patients belonged to the good prognosis group. The second group of patients, the test set, consists of 19 patients of which 12 patients belonged to the poor prognosis group and 7 patients belonged to the good prognosis group. The microarray data was already background corrected, normalized and log-transformed. Preprocessing step removes genes with small profile variance, less than the 10th percentile.

The last data sets consists of transcript profiles of 251 primary breast tumors were assessed by using Affymetrix U133 oligonucleotide microarrays. cDNA sequence analysis revealed that 58 of these tumors had p53 mutations resulting in protein-level changes, whereas the remaining 193 tumors were p53 wt [14].

2.2 Clinical Data

The first data of 129 patients contained information on 17 available clinical variables, 5 were excluded [11]: two redundant variables that were least informative based on univariate analysis in those variable pairs with a correlation coefficient exceeding 0.7, and three variables with too many missing values. After exclusion of patients with missing clinical information, this data set consisted of 110 patients remained in 85 of whom disease did not recur whilst in 25 patients disease recurred.

The second data in which response to treatment was studied, consisted of 12 variables for 133 patients [12]. Patient and variable exclusion as described above resulted in this data set. Of the 129 remaining patients, 33 showed complete response to treatment while 96 patients were characterized as having residual disease.

In the third data, relapse was studied in 187 patients [13]. After preprocessing, this data set retained information on 5 variables for 177 patients. In 112 patients, no relapse occurred while 65 patients were characterized as having a relapse.

The fourth data [3] consisted of predefined training and test sets same as that of corresponding microarray data. The last data set consisted of 251 patients with 6 available clinical variables [14]. After exclusion of patients with missing clinical information, this data set consisted of 237 patients of which 55 patients with p53 mutant breast tumor and the remaining patients without p53 mutant breast tumor.

3 Methods

In the first section, we will discuss about GEVD and represent it in terms of ordinary EVD. Then an overview of LS-SVM formulation to kernel PCA and least squares support vector machines (LS-SVM) will be given. Next, we formulate an optimization problem for kernel GEVD in primal space and solution in dual space. Finally, by generalizing this optimization problem in terms of LS-SVM classifier, we propose a new machine learning approach for data fusion and classifications, a weighted LS-SVM classifier.

3.1 Generalized Eigenvalue Decomposition

The Generalized Singular Value Decomposition (GSVD) of $m \times N$ matrix A and $p \times N$ matrix B is [16]

$$A = U\Sigma_A X^T \tag{1}$$

$$B = V\Sigma_B X^T \tag{2}$$

where U, V are orthogonal matrices and columns of X are generalized singular vectors.

If B^TB is invertible, then the GEVD of A^TA and B^TB can be obtained from Equations (1) and (2) as follows:

$$A^TA(X^T)^{-1} = B^TB(X^T)^{-1}\Lambda. \tag{3}$$

where Λ is a diagonal matrix with diagonal entries $\Lambda_{ii} = (\frac{\Sigma_{A_{ii}}}{\Sigma_{B_{ii}}})^2$, $i = 1, \ldots, N$.

Equation (3) can be represented in eigenvalue decomposition (EVD) as follows:

$$(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}U = U\Lambda.$$

where $U = (B^TB)^{1/2}(X^T)^{-1}$. The SVD of matrix $A(B^TB)^{-1/2}$ is given below:

$$A(B^TB)^{-1/2} = V\Lambda U^T. \tag{4}$$

The matrix $(B^TB)^{-1/2}$ is defined [17] as follows: Let EVD of $B^TB = T\Sigma T^T$, where columns of T are eigenvectors and Σ is a diagonal matrix. $(B^TB)^{1/2} = T\Sigma^{1/2}T^T$ and $(B^TB)^{-1/2} = TQT^T$, where Q is a diagonal matrix with diagonal entries $Q_{ii} = (\Sigma_{ii})^{-1/2}$, $i = 1, \ldots, N$.

3.2 LS-SVM formulation to Kernel PCA

An LS-SVM approach to kernel PCA was introduced in [18]. This approach showed that kernel PCA is the dual solution to a primal optimization problem formulated in a kernel induced feature space. Given training set $\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$, the LS-SVM approach to kernel PCA is formulated in the primal as:

$$\min_{w,e_i,b} J_p(w, e_i) = \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2 - \frac{1}{2}w^Tw,$$

such that $e_i = w^T\varphi(x_i) + b$, $i = 1, \ldots, N$, where b is a bias term and $\varphi(.)$: $\mathbb{R}^d \to \mathbb{R}^{d_h}$ is the feature map which maps the d-dimensional input vector x from the input space to the d_h-dimensional feature space.

Kernel PCA in dual space takes the form:

$$\Omega_c\alpha = \lambda\alpha$$

where α is an eigenvector, λ an eigenvalue and Ω_c denotes the centered kernel matrix with ijth entry: $\Omega_{c,i,j} = K(x_i, x_j)$ - $\frac{1}{N}\sum_{r=1}^{N} K(x_i, x_r)$ - $\frac{1}{N}\sum_{r=1}^{N} K(x_j, x_r)$ + $\frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} K(x_r, x_s)$, with $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ a positive definite kernel function.

3.3 Least squares support vector machine classifiers

A kernel algorithm for supervised classification is the SVM developed by Vapnik [19] and others. Contrary to most other classification methods and due to the way data are represented through kernels, SVMs can tackle high dimensional data (for example microarray data). Given a training set $\{x_i, y_i\}_{i=1}^N$ with input data $x_i \in \mathbb{R}^d$ and corresponding binary class labels $y_i \in \{-1, +1\}$, the SVM forms a linear discriminant boundary $y(x) = \text{sign}[w^T\varphi(x) + b]$ in the feature space with maximum distance between samples of the two considered classes, with w representing the weights for the data items in the feature space, b the bias term and $\varphi(.)$: $\mathbb{R}^d \to \mathbb{R}^{n_1}$ is the feature map which maps the d-dimensional input vector x from the input space to the n_1-dimensional feature space. This corresponds to a non-linear discriminant function in the original input space. Vapnik's SVM classifier formulation was modified in [20]. This modified version is much faster for classification because a linear system instead of a quadratic programming problem needs to be solved.

The constrained optimization problem for least squares support vector machine (LS-SVM) [20, 21] for classification are defined as follows:

$$\min_{w,b,e} \frac{1}{2}w^T w + \gamma\frac{1}{2}\Sigma_{i=1}^N e_i^2$$

subject to:

$$y_i[w^T\varphi(x_i) + b] = 1 - e_i, \qquad i = 1, \ldots, N$$

with e_i the error variables, tolerating misclassifications in cases of overlapping distributions, and γ the regularization parameter, which allows tackling the problem of overfitting. The LS-SVM classifier formulation implicitly correspond to a regression interpretation with binary target y_i=$\pm$1.

In the dual space the solution is given by

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + \frac{I}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix}$$

with $y = [y_1, \ldots, y_N]^T$, $1_N = [1, \ldots, 1]^T$, $e = [e_1, \ldots, e_N]^T$, $\beta = [\beta_1, \ldots, \beta_N]^T$, $\Omega_{i,j} = y_i y_j K(x_i, x_j)$ where $K(x_i, x_j)$ is the kernel function.

The classifier in the dual space takes the form

$$y(x) = \text{sign}[\sum_{i=1}^N \beta_i y_i K(x, x_i) + b]$$

where β_i are Lagrange multipliers.

3.4 LS-SVM and kernel GEVD

LS-SVM formulations to different problems were discussed in [21]. This class of kernel machines emphasizes primal-dual interpretations in the context of constrained optimization problems. In this section we discuss LS-SVM formulations to kernel GEVD, which is a non-linear GEVD of $m \times N$ matrix A, and $p \times N$ matrix B, and a weighted LS-SVM classifier.

Given a training data set of N points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^N$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$ ($x_i^{(1)}$ and $x_i^{(2)}$ are the i^{th} sample of matrices A and B respectively).

Consider the feature maps $\varphi^{(1)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_1}$ and $\varphi^{(2)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_2}$ to a high dimensional feature space $\mathcal{F}$, which is possibly infinite dimensional. The centered feature matrices $\Phi_c^{(1)} \in \mathbb{R}^{n_1 \times N}$, $\Phi_c^{(2)} \in \mathbb{R}^{n_2 \times N}$ become

$$\Phi_c^{(1)} = [\varphi^{(1)}(x_1^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T; \ldots; \varphi^{(1)}(x_N^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T]^T$$

$$\Phi_c^{(2)} = [\varphi^{(2)}(x_1^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T; \ldots; \varphi^{(2)}(x_N^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T]^T,$$

where $\hat{\mu}_{\varphi l} = \frac{1}{N}\Sigma_{i=1}^N \varphi^{(l)}(x_i^{(l)})$, $l = 1, 2$

3.4.1 LS-SVM approach to Kernel GEVD

Kernel GEVD is a nonlinear extension of GEVD, in which data are first embedded into a high dimensional feature space introduced by kernel and then linear GEVD is applied. While considering the matrix $A(B^TB)^{-1/2}$ in Equation (4) and the feature maps $\varphi^{(1)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_1}$ and $\varphi^{(2)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_2}$ described in previous section, the covariance matrix of $A(B^TB)^{-1/2}$ in the feature space becomes $C \approx \Phi_c^{(1)}(\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1}\Phi_c^{(1)^T}$ with eigendecomposition $Cv = \lambda v$.

While considering kernel PCA formulation based on the LS-SVM framework [22] was discussed in section 3.2 and EVD of $Cv = \lambda v$ in primal space, our objective is to find the directions in which projected variables have maximal variance.

The LS-SVM approach to kernel GEVD is formulated as follows:

$$\boxed{\begin{array}{c} \min_{v,e} \; J(v,e) \; = \; \gamma \frac{1}{2} e^T (\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1} e - \frac{1}{2} v^T v \\ \\ \textit{such that} \quad e = \Phi_c^{(1)^T} v \; , \end{array}} \tag{5}$$

where v is the eigenvector in the primal space, $\gamma \in \mathbb{R}^+$ is a regularization constant and e are the projected data points to the target space.

Defining the Lagrangian

$\mathcal{L}(v, e; \alpha) = \frac{\gamma}{2} e^T e - \frac{1}{2} v^T v - \alpha^T \{(e - \Phi_c^{(1)^T} v)\}$,

with optimality conditions,

$\frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow v = \Phi_c^{(1)} \alpha$,

$\frac{\partial \mathcal{L}}{\partial e} = 0 \rightarrow \alpha = \gamma (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e$,

$\frac{\partial \mathcal{L}}{\partial \alpha} = 0 \rightarrow e = \Phi_c^{(1)^T} v$,

elimination of v and e will yield an equation in the form of GEVD

$$\boxed{\Omega_c^{(1)} \alpha = \lambda \Omega_c^{(2)} \alpha,}$$

where $\lambda = \frac{1}{\gamma}$ largest eigenvalue, $\Omega_c^{(1)}$, $\Omega_c^{(2)}$ are centered kernel matrices and α are generalized eigenvectors. The symmetric kernel matrices $\Omega_c^{(1)}$ and $\Omega_c^{(2)}$ resolves the heterogeneities of clinical and microarray data by the use of kernel trick, where data which have diverse data structures are transformed into kernel matrices with same size.

In a special case of GEVD, if one of the data matrix is identity matrix, it will be equivalent to ordinary EVD. If $(\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} = I$, then the optimization problem proposed for kernel GEVD (See Equation (5)) will be equivalent to optimization problem in [18] for the LS-SVM approach to kernel PCA.

3.4.2 Weighted LS-SVM classifier

Our objective is to represent kernel GEVD in the form of weighted LS-SVM classifier. Given the link between LS-SVM approach to kernel GEVD in Equation (5) and the weighted LS-SVM classifier (see [23] in a different type of weighting to achieve robustness), one considers the following optimization problem in primal weight space:

$$\boxed{\begin{array}{c} \min_{v,e,b} \; J(v,e) \; = \; \gamma \frac{1}{2} e^T (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e + \frac{1}{2} v^T v \\ \\ \textit{such that} \quad y = \Phi_c^{(1)^T} v + b 1_N + e, \end{array}}$$

with $e = [e_1, \ldots, e_N]^T$ a vector of variables to tolerate misclassifications, weight vector v in primal weight space, bias term b and regularization parameter $\gamma \in \mathbb{R}^+$. Compared to the constrained optimization problem for least squares support vector machine (LS-SVM) [20, 21], in this case, the error variables are weighted with a matrix $(\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1/2}$.

The weight vector v can be infinite dimensional, which makes the calculation of v impossible in general. One defines the Lagrangian

$\mathcal{L}(v, e, b; \alpha) = \frac{1}{2} v^T v + \frac{\gamma}{2} e^T (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e - \alpha^T \{((\Phi_c^{(1)^T} v) + b 1_N) + e - y\},$

with Lagrange multipliers $\alpha \in \mathbb{R}^N$.

$\frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow v = \Phi_c^{(1)} \ \alpha$

$\frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow 1_N^T \alpha = 0$

$\frac{\partial \mathcal{L}}{\partial e} = 0 \rightarrow \alpha = \gamma (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e$

$\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \rightarrow e + \Phi_c^{(1)^T} v + b = y$

Elimination of v and e yields a linear system

$$\begin{bmatrix} 0 & 1_N^T \\ 1_N & \Omega_c^{(1)} + \gamma^{-1} \Omega_c^{(2)} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{6}$$

with $y = [y_1, ..., y_N]^T$, $1_N = [1, ..., 1]^T$, $\alpha = [\alpha_1, ..., \alpha_N]^T$, $\Omega_c^{(1)} = \Phi_c^{(1)^T} \Phi_c^{(1)}$ and $\Omega_c^{(2)} = \Phi_c^{(2)^T} \Phi_c^{(2)}$.

The resulting classifier in the dual space is given by

$$y(x) = \sum_{i=1}^{N} \alpha_i ([K^{(1)}(x, x_i) + \frac{1}{\gamma} K^{(2)}(x, x_i)] + b) \tag{7}$$

with α_i are the Lagrange multipliers, γ is a regularization parameter has chosen by user, $K^{(1)}(x, z) = \varphi^{(1)}(x)^T \varphi^{(1)}(z)$, $K^{(2)}(x, z) = \varphi^{(2)}(x)^T \varphi^{(2)}(z)$ and $y(x)$ is the output corresponding to validation point x. The LS-SVM for nonlinear function estimation in [23] is similar to the proposed weighted LS-SVM classifier.

The symmetric, kernel matrices $K^{(1)}$ and $K^{(2)}$ resolve the heterogeneities of clinical and microarray data sources such that they can be merged additively as a single kernel. The optimization algorithm for the weighted LS-SVM classifier is given below:

Algorithm: Optimization algorithm for the weighted LS-SVM classifier

1. Given a training data set of N points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^N$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$.

2. Calculate Leave-One-Out cross validation (LOO-CV) performances of training set with different combinations of γ and σ_1, σ_2 (bandwidths of kernel functions $K^{(1)}$, $K^{(2)}$) by solving linear system Equation (6) and Equation (7). In case the Leave-One-Out (LOO) approach is computationally expensive, one could replace it with a leave p group out strategy (p-fold cross-validation)

3. Obtain the optimal parameters combinations (γ, σ_1, σ_2) which have the highest LOO-CV performance.

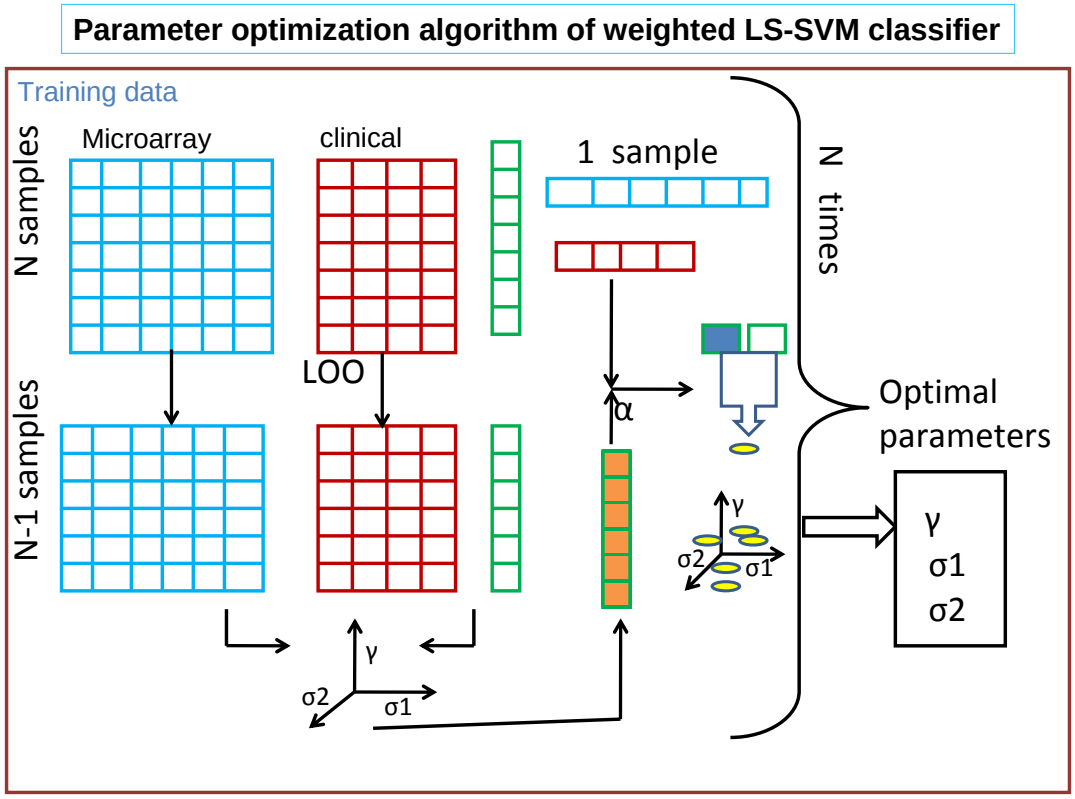


Figure 1: Overview of algorithm. The data sets represented as matrices with rows corresponding to patients and columns corresponding to genes and clinical parameters respectively for first and second data sets. LOO-CV is applied to select the optimal parameters.

The proposed optimization problem is similar to the the weighted LS-SVM formulation in [22] which replaced $(\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1}$ with a diagonal matrix to achieve sparseness and robustness.

The proposed method is a new machine learning approach in data fusion and subsequent classifications. In this study, the advantages of a weighted LS-SVM classifier were explored, by designing a clinical classifier. This clinical classifier combined kernels by weighting kernel inner product from one data set with that from the other data set. Here we considered microarray kernels as weighting matrix for clinical kernels. In each of these case studies, we compared the prediction performance of individual data sets with GEVD, kernel GEVD and weighted LS-SVM classifier. In kernel GEVD, σ_1 and σ_2 are the bandwidth of RBF-kernel function $K(x,z) = \exp(-\frac{||x-z||^2}{2\sigma^2})$ of clinical and microarray data sets respectively. These parameters were chosen such that the pairs (σ_1, σ_2) which obtained the highest LOO-CV performance. The parameter selection (see Algorithm) for the weighted LS-SVM classifier

are illustrated in Figure 1. For several possible values of the kernel parameters σ_1 and σ_2, the LOO cross validation performance is computed for each possible combinations of γ. The optimal parameters are the combinations (σ_1, σ_2, γ) with best LOO-CV performance. Remark the complexity of this optimization procedure because both the kernel parameters (σ_1 and σ_2) and γ need to be optimized in the sense of the LOO-CV performance.

4 Results

In all case studies except fourth, 2/3rd of the data samples of each class are assigned randomly to the training and the rest to the test set. These randomization are the same for all numerical experiments on all data sets. This split was performed stratified 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. In all these cases, the microarray data were standardized to zero mean and unit variance. Normalization of training sets as well as test sets are done by using the mean and standard deviation of each gene expression profile of the training sets. In the fourth data set [3], all data samples have already been assigned to a training set or test set.

Initially LS-SVM classifiers have been applied on individual data sets: clinical and microarray. Then we performed GEVD on training samples of clinical and microarray data sets and obtained generalized eigenvectors (GEV). Scores are obtained by projecting the clinical data on to the directions of GEV. LS-SVM model is trained and validated on scores corresponding to training set and test set respectively.

4.1 Kernel GEVD

The optimal parameters of the kernel GEVD (bandwidths of clinical and microarray kernels) are selected using LOO-CV performance. We applied kernel GEVD on microarray and clinical kernels. Then we obtained the scores by projecting clinical kernels on to the direction of kernel GEV. Similar to GEVD, LS-SVM model is trained and validated on scores corresponding to training set and test set respectively. High-throughput data such as microarray have used only for the model development. The results show that considerations of two data sets in a single framework improve the prediction performance than individual data sets. In addition, kernel GEVD significantly improve the classification performance over GEVD. The results of the five case studies are shown in Table 2 and Figure 2. We represent expression and clinical data with kernel matrix, based on RBF kernel function. The RBF kernel functions makes each of the these data which has diverse structures, transformed into kernel matrices with same size.

4.2 Weighted LS-SVM classifier

We proposed a weighted LS-SVM classifier, a useful technique in data fusion as well as in supervised learning. The parameters (γ in Equation (6) and σ_1, σ_2 the bandwidths of microarray and clinical kernel functions) associated with this method are selected by Algorithm. In each LOO-CV, 1 - samples are left out and models are built for all possible combinations of parameters on the remaining $N-1$ samples. The optimization problem is not sensitive to small changes of bandwidths of microarray and clinical kernel functions. Careful tuning of γ allows tackling the problem of overfitting and tolerating misclassifications. Models parameters are chosen corresponding to the model with highest LOO AUC. The LOO-CV approach takes less than a minute for a single iteration of the first three case studies and 1-2 minutes for the rest of case studies. Statistical significance test are performed in order to allow a correct interpretations of the results. A non-parametric paired test, the Wilcoxon signed rank test (signrank in Matlb) [24], has been used in order to make general conclusions. A threshold of 0.05 is respected, which means that two results are significantly different if the value of the Wilcoxon signed rank test applied to both of them is lower than 0.05. On all case studies, weighted LS-SVM classifier outperformed all other discussed methods, in terms of test AUC, as shown in Table 2 and Figure 2. The weighted LS-SVM performance on second and fourth cases slightly better, but not significantly, than the kernel GEVD.

To compare LS-SVM with other classification methods, we have applied Naive Bayes classifiers individually to clinical and microarray data. In this case, the normal distribution were used to model continuous variables, while ordinal and nominal variables were modeled with a multivariate multinomial distribution. The average test AUC of this method, when applied on five case studies are shown in Table 4.

Then we compare the proposed weighted LS-SVM classifiers with other data fusion techniques which integrate microarray and clinical data sets. Daemen *et al* [5] investigated the effect of data integration on performance with three case studies [11–13]. They reported that a better performance was obtained when considering both clinical and microarray data with the weights (μ) assigned to them optimized (μClinical+(1-μMicroarray)). In addition they concluded from their 10-fold AUC measurements that the clinical kernel variant, led to a significant increase in performance, in the kernel based integration approach of clinical and microarray. The first three case studies, we have taken from the work of Daemen *et al* [5]. They have considered the 200 most differential genes selected from the training data with the Wilcoxon rank sum test, for the kernel matrix obtained from microarray. The fourth case study, we have taken from the paper of Gevaert *et al* [2] in which they investigated different types of integration strategies, with Bayesian network classifier. They concluded that partial integration performs better in terms of test AUC. Our results also confirms that consideration of microarray and clinical data sets together, improves prediction performances than individual data sets.

In our analysis, microarray-based kernel matrix are calculated on large data set without preselecting genes and thus avoiding potential selection bias [25]. In addition, we compared RBF kernel with the clinical kernel function [5] in terms of LOO-CV performance. Results are given on Table 3. We followed the same strategy which was explained for weighted LS-SVM classifier, except the clinical kernel function have been used for the clinical parameters. On three out of five case studies, RBF kernel functions performs better than clinical kernel function.

5 Discussion

Integrative analysis has been primarily used to prioritize disease genes or chromosomal regions for experimental testing, to discover disease subtypes or to predict patient survival or other clinical variables. The ultimate goal of this work is to propose a machine learning approach which is functional in both data fusion and supervised learning. We further analyzed the potential benefits of merging microarray and clinical data sets for prognostic application in breast cancer diagnosis.

We integrate microarray and clinical data into one mathematical model, for the development of highly homogeneous classifiers in clinical decision support. For this purpose, we present a kernel based integration framework in which each data set is transformed into a kernel matrix. Integration occurs on this kernel level without referring back to the data. Some studies [1,5] already reported that intermediate integration of clinical and microarray data sets improves prediction performance on breast cancer outcome. In primal space, the clinical classifier is weighted with expression values. The solution in dual space is given on Equations (6) and (7) which provides a way to integrate two kernel functions explicitly and perform further classifications.

To verify the merit of the proposed approach over the single data sources such as clinical and microarray data, the LS-SVM were built on all data sets individually for classifying cancer patients. Next, GEVD and kernel GEVD are performed. Then the projected variances in the new space (scores) have used to build the LS-SVM. Finally weighted LS-SVM approach was used for the integration of both microarray and clinical kernel functions and performed subsequent classifications. Thus weighted LS-SVM classifier proposes a new optimization framework to solve the problem of classification using features of different types such as clinical and microarray data.

We should note that the models proposed in this paper are expensive, but less than the other kernel-based data fusion techniques. Since the proposed weighted LS-SVM classifier simplified both data fusion and classification in a single framework, it does not have an additional cost for tuning parameters for kernel-based classifiers. And it is given that, the weighting matrix should be invertible in the optimization problem of kernel GEVD and the weighted LS-SVM classifier.

In life science research, there is an increasing need for heterogeneous data integration such as proteomics,

genomics, mass spectral imaging and so on. Such studies are required to determine, which data sets are most significant to be considered as weighting matrix. The proposed weighted LS-SVM classifier integrates heterogeneous data sets to achieve good performing and affordable classifiers.

6 Conclusion

The results suggest that the use of our integration approach on gene expression and clinical data can improve the performance of decision making in cancer. We proposed a weighted LS-SVM classifier for the integration of two data sources and further prediction task. Each data set is represented with a kernel matrix, based on the RBF kernel function. The proposed clinical classifier gives a step towards improving predictions for individual patients about prognosis, metastatic phenotype and therapy responses.

Because the parameters (bandwidth for kernel matrices and regularization term γ of weighted LS-SVM) had to be optimized, all possible combinations of these parameters were investigated with a LOO-CV. Since these parameters optimization strategy is time consuming, one can further investigate a parameter optimization criterion for kernel GEVD and weighted LS-SVM.

The applications of proposed method are not limited to clinical and expression data sets. Possible additional applications of weighted LS-SVM include integration of genomic information collected from different sources and biological processes. In short, the proposed machine learning approach is a promising mathematical framework in both data fusion and non-linear classification problems.

7 Competing interests

The authors declare that they have no competing interests.

8 Authors' contributions

MT performed the kernel based data integration modeling and drafted the paper. KDB and JS participated in the design and implementation of framework. KDB, JS and BDM helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

BDM is full professor at the Katholieke Universiteit Leuven, Belgium. Johan AK Suykens is professor at the Katholieke Universiteit Leuven, Belgium. Research supported by: Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, START 1, OT 09/052 Biomarker, several PhD/postdoc fellow grants;

Industrial Research fund (IOF): IOF/HB/13/027 Logic Insulin, IOF: HB/12/022 Endometriosis; Flemish Government: FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits), research community MLDM; IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping, TBM Rectal Cancer, TBM IETA; Hercules Stichting: Hercules III PacBio RS;iMinds: SBO 2013; ArtD Instance;IMEC: phd grant;VLK van der Schueren: rectal cancer;VSC Tier 1: exome sequencing; Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate); COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network; CoE EF/05/006, IUAP DYSCO, FWO G.0377.12, ERC AdG A-DATADRIVE-B. The scientific responsibility is assumed by its authors.

References

1. van Vliet MH, Horlings HM, van de Vijver M, Reinders MJT: **Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome**. *PLoS ONE* 2012, **7**:e40385–e40358.
2. Gevaert O, Smet F, Timmerman D, Moreau Y, De Moor B: **Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks**. *Bioinformatics* 2006, **22**:e184–e190.
3. van't Veer LJ, Dai H, Van De Vijver MJ, HeY D, Hart AAM, Mao M, Peterse HL, Van Der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernard R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**:530–536.
4. Boulesteix AL, Porzelius C, Daumer M: **Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value**. *Bioinformatics* 2008, **24**:1698–1706.
5. Daemen A, Timmerman D, Van den Bosch T, Bottomley C, Kirk E, Van Holsbeke C, Valentin L, Bourne T, De Moor B: **Improved modeling of clinical data with kernel methods**. *Artificial Intelligence in Medicine* 2012, **54**:103–114.
6. Pittman J, Huang E, Dressman H, Horng C, Cheng S, Tsou M, Chen C, Bild A, Iversen E, Huang A, Nevins J, West M: **Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes**. *PNAS* 2004, **101**:8431–8436.
7. Sedeh S R, Bathe M, K-J B: **The Subspace Iteration Method in Protein Normal Mode Analysis**. *J Comput Chem* 2010, **31**:66–74.
8. Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genomescale expression data sets of two different organisms**. *PNAS* 2003, **100**:3351–3356.
9. Chu F, Wang L: **Application of Support Vector Machine to Cancer Classification with Microarray Data**. *International Journal of Neural systems, World Scientific* 2005, **5**:475–484.
10. Chun LH, Wen CL: **Detecting differentially expressed genes in heterogeneous disease using half Student's t-test**. *Int.I.Epidemiol* 2010, **10**:1–8.
11. Chin K, De Vries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies**. *Cancer Cell.* 2006, **10**:529–541.
12. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, Rouzier R, Sneige N, Ross JS, Vidaurre T, Gómez HL, Hortobagyi GN, Pusztai L: **Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer**. *J Clin Oncol* 2006, **24**:4236–4244.
13. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, van de Vijver MJ, Bergh J, Piccart M, Delorenzi M: **Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis**. *J Natl Cancer Inst* 2006, **98**:262–272.
14. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival**. *PNAS* 2005, **102**(38):13550–13555.
15. Dai M, Wang AD Pand Boyd, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson F S Jand Meng: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res.* 2005, **33**:e175.
16. Golub GH, Van Loan CF: *Matrix Computations*. 2nd ed. (Baltimore: Johns Hopkins University Press) 1989.
17. Higham N: **Newton's Method for the Matrix Square Root**. *Mathematics of Computation* 1986, **46**(174):537–549.
18. Suykens JAK, Van Gestel T, Vandewalle J, De Moor B: **A support vector machine formulation to PCA analysis and its kernel version**. *IEEE Transactions on Neural Networks* 2003, **14**(2):447–450.
19. Vapnik V: *The Nature of Statistical Learning Theory*. Springer-Verlag 1995.
20. Suykens JAK, Vandewalle J: **Least Squares Support Vector Machine classifiers**. *Neural Processing Letters* 1999, **9**:293–300.

21. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J: *Least Squares Support Vector Machines*. Singapore: World Scientific 2002.
22. Alzate C, Suykens JAK: **Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2010, **32**(2):335–347.
23. Suykens JAK, De Brabanter J, Lukas L, Vandewalle J: **Weighted least squares support vector machines: robustness and sparse aprroximation**. *Neurocomputing* 2002, **48**:85–105.
24. Dawson-Saunders B, Trapp RG: *Basic Clinical Biostatistics*. Prentice-Hall International Inc., London 1994.
25. Ambroise C, McLachlan GJ: **Selection bias in gene extraction on the basis of microarray gene-expression data**. *PNAS* 2002, **99**:6562–6566.

Figures

1. Overview of algorithm.

The data sets represented as matrices with rows corresponding to patients and columns corresponding to genes, clinical parameters for first and second data sets respectively. Leave-One-Out cross validation is used to select the optimal parameters.

2. Comparison of the prediction accuracy of the classifiers.

Boxplots of the test AUC values obtained in 100 repetitions for 5 breast cancer cases.(a) Case I (b) Case II (c) Case III (d) Case IV (e) Case V.

Tables

Table 1: Summary of the preprocessed 5 breast cancer data sets.

Table 1: Summary of the 5 breast cancer data sets.

Case Study	#Samples		#Genes	#Clinical variables
	Class 1	Class2		
Case I	85	25	5000	Age, Ethnicity, ER status, PR status, Radiation treatment, Chemotherapy, Hormonal therapy, Nodal status, Metastasis, Tumor stage, Tumor size, Tumor grade.
Case II	33	96	6000	Age, Ethnicity, pretreatment tumor stage, nodal status, nuclear grade, ER status, PR status, HER2 status.
Case III	112	65	5000	Age, Tumor size, Nodal status, ER status, Tamoxifen treatment.
Case IV	46	51	12192	Age, Tumor size, Grade, Erp, Angioinvasion, Lymphocytic Infiltrate, PRp.
Case V	58	193	20055	Age, Tumor size, Grade, ER, Prp, Lymph node.

Table 2. Comparisons of different classifiers : test AUC(std) of breast cancer cases
Table 3. Comparisons of RBF with clinical kernel functions
Table 3. Naive Bayes classifiers performance on clinical and microarray data sets

Table 2: Comparisons of different classifiers : test AUC(std) of breast cancer cases

Classifiers	Case I	Case II	Case III	Case IV	Case V
CL +LS-SVM					
test AUC	0.7795(0.0687)	0.7772(0.0554)	0.6152(0.0565)	0.6622(0.0628)	0.7740(0.0833)
p-value	0.0039	1.48E-04	0.0086	5.21E-06	0.1602
MA+LS-SVM					
test AUC	0.7001(0.0559)	0.8065(0.0730)	0.6217(0.0349)	0.7357(0.0085)	0.6166(0.0508)
p-value	0.0059	0.0140	0.0254	2.41E-04	0.0020
GEVD+LS-SVM					
test AUC	0.7801(0.0717)	0.7673(0.0548)	0.6196(0.0829)	0.7730(0.1011)	0.8001(0.0648)
p-value	0.0137	3.41E-05	0.0040	0.1558	0.0840
KGEVD+LS-SVM					
test AUC	0.7982(0.0927)	0.8210(0.0670)	0.6437(0.0313)	0.7901(0.0917)	0.8031(0.0624)
p-value	0.0195	0.1144	0.0020	0.6162	0.0720
weighted LS-SVM					
test AUC	**0.8177**(0.0666)	**0.8465**(0.0480)	**0.6985**(0.0443)	**0.8119**(0.0893)	**0.8210**(0.0477)

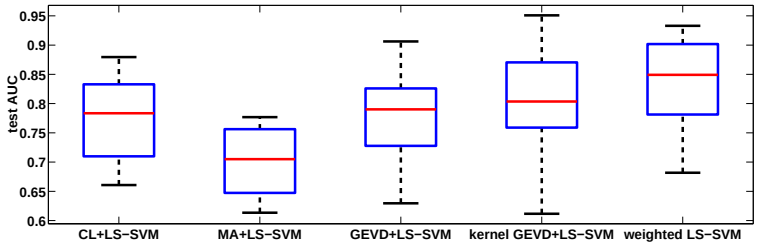
p-value: a paired test, Wilcoxon signed rank test.
CL and MA are the clinical and microarray kernels of RBF kernel functions.

Table 3: Comparisons of RBF with clinical kernel functions in terms of LOO-CV performances

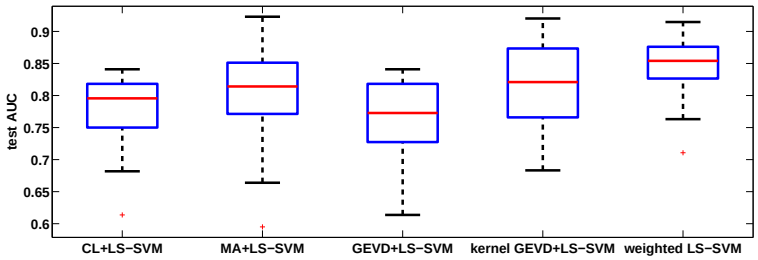
Kernel functions	Case I	Case II	Case III	Case IV	Case V
Clinical kernel	0.8108(0.0351)	**0.8315** (0.0351)	**0.7479**(0.0111)	0.7385(0.1100)	0.7673(0.0213)
RBF	**0.8243**(0.0210)	0.8202(0.0100)	0.7143(0.0217)	**0.7846**(0.0699)	**0.7862**(0.0221)

Table 4: Naive Bayes classifiers performance on clinical and microarray data sets in terms of test AUC(std)

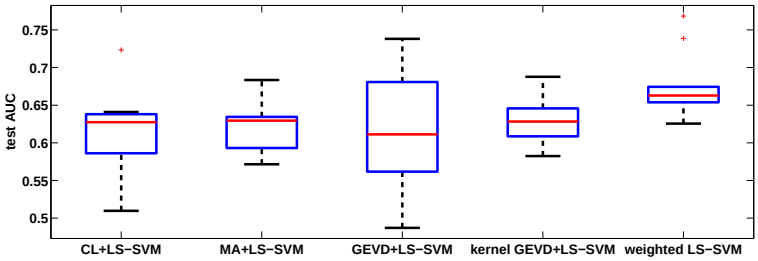
Data Source	Case I	Case II	Case III	Case IV	Case V
Clinical Data	0.6235(0.0912)	0.739 (0.0722)	0.5533(0.0438)	0.7156(0.0503)	0.6767(0.0513)
Microarray	0.5028(0.037)	0.6662(0.088)	0.5324(0.0616)	0.6011(0.0699)	0.5189(0.0412)



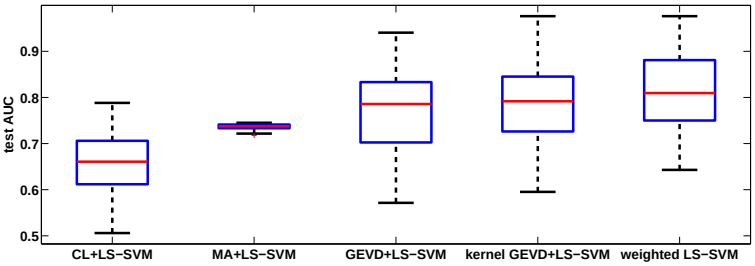
(a)



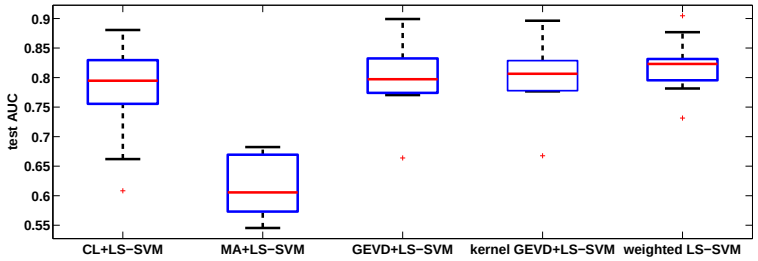
(b)



(c)



(d)



(e)

Figure 2: Boxplots of the test AUC values obtained in 100 repetitions for 5 breast cancer cases.(a) Case I (b) Case II (c) Case III (d) Case IV (e) Case V.