Deepening the methodology behind Data integration and Dimensionality reduction:

Applications in Life Sciences

Thomas, Minta

ESAT, KU Leuven, Belgium

April 25, 2017

1 – Outline

1. Chapter 1: Introduction
2. Chapter 2: Maximum Likelihood Estimation via Generalized Eigenvectors
3. Chapter 3: Robust PCA
4. Chapter 4: New Bandwidth Selection Criterion for KPCA
5. Chapter 5: Weighted LS-SVM Classifier
6. Chapter 6: Chemical Descriptor based on Connection-Table
7. Chapter 7:Conclusion and Future Studies

1 – Introduction

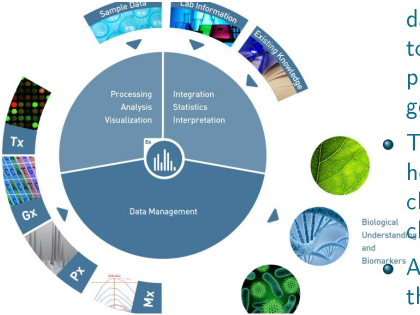
1-1 General Introduction

- 1-1-1 Motivation
- 1-1-2 Data Integration
- 1-1-3 Dimensionality Reduction
- 1-1-4 Applications in Life Sciences

1-2 Methodology: Overview

- 1-2-1 Dimensionality Reduction Techniques
- 1-2-2 Kernel Methods
- 1-2-3 LS-SVM Classifier
- 1-2-4 Performance Measures

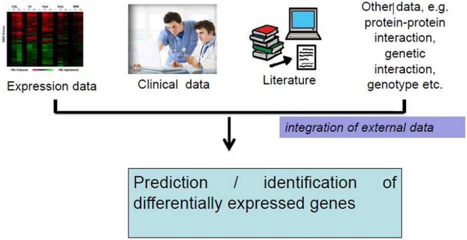
1 – 1-1 Motivation



- Data integration plays important roles in combining clinical, and environmental data with high-throughput genomic data to identify functions of genes, proteins,and other aspects of the genome.
- The problems of high dimensionality and heterogeneity of data always raise lots of challenges in computational biology and chemistry.
- As the size of data sets increase, as well their complexity, dimensionality reduction and advanced analytics will gain importance.

1 – 1-2 Data Integration

- Data integration can be defined as the process of combining data residing in diverse sources to provide users with a comprehensive view of such data



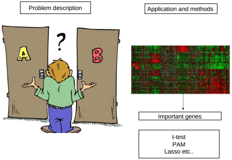
Expression data
Clinical data
Literature
Other data, e.g. protein-protein interaction, genetic interaction, genotype etc.
integration of external data
Prediction / identification of differentially expressed genes

Applications

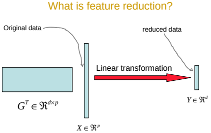
- Tumor classification: A reliable and precise classification of tumors is essential for successful treatment of cancer
- The identification of “marker” genes that characterize the different tumor classes.

1 – 1-3 Dimensionality Reduction

- Feature Selection also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.



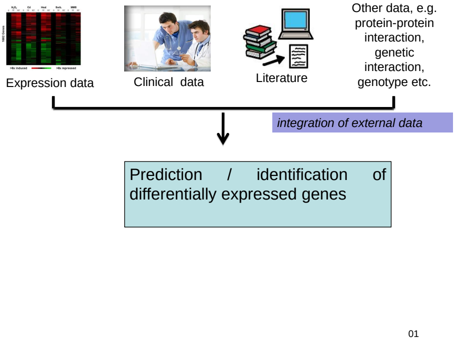
- Feature Extraction transforms the data in the high-dimensional space to a space of fewer dimensions.



$$G \in \Re^{p \times d} : X \rightarrow Y = G^T X \in \Re^d$$

1 – 1-4 Application in Life Sciences

- Microarray Data Analysis
- Quantitative structure–activity relationship models
- Clinical Data Analysis



Expression data

Clinical data

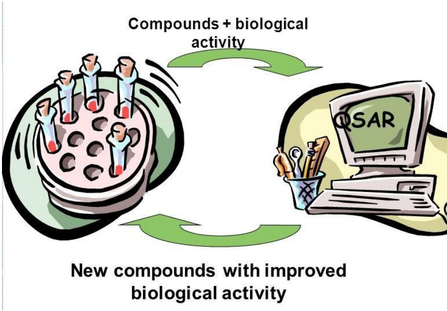
Literature

Other data, e.g.
protein-protein
interaction,
genetic
interaction,
genotype etc.

integration of external data

Prediction / identification of differentially expressed genes

01



Compounds + biological activity

New compounds with improved biological activity

If we have an $m \times n$ matrix A,

Singular Value Decomposition (SVD) of A

$$A = USV^T,$$

where S is a diagonal matrix and U and V are orthogonal matrices.

Eigenvalue Decomposition (EVD) of A

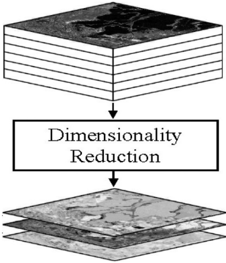
$$A^T A = VLV^T,$$

where L is a diagonal matrix and V is orthogonal $n \times n$ matrix.

Principal Component Decomposition of A

$$T = AV,$$

where V is an $n \times n$ matrix whose columns are the eigenvectors of $A^T A$.



If we have an $m \times n$ matrix A and $p \times n$ matrix B

Generalized Singular Value Decomposition (GSVD) of A and B

$$A = U\Sigma_A X^T$$

$$B = V\Sigma_B X^T$$

where U, V are orthogonal matrices, the columns of X are generalized singular vectors and Σ_A, Σ_B are diagonal matrices

Generalized Eigenvalue Decomposition (GEVD) of the matrix pair $A^T A$, $B^T B$,

$$A^T A(X^T)^{-1} = B^T B(X^T)^{-1}\Lambda$$

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

- **Kernel Methods** are class of algorithms for pattern analysis which find and study general types of relations in a data sets.

The functions that are most frequently employed in classification problems are

- **Linear Kernel**: ${x^i}^T x^j$
- **Polynomial Kernel**: $({x^i}^T x^j + b)^d$ with - as kernel parameters - the intercept constant $b \in \mathbb{R}^+$ and degree $d \in \mathbb{N}$
- **Radial Basis Function (RBF)**: $\exp(-||x^i - x^j||_2^2/\sigma^2)$ with $\sigma \in \mathbb{R}^+$ representing the bandwidth.

1 – 2-3 LS-SVM Classifier

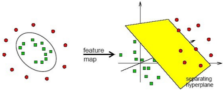
The constrained optimization problem for an LS-SVM for classification has the following form:

$$\boxed{\min_{w,b,e}(\frac{1}{2}w^T w + \gamma\frac{1}{2}\Sigma_{k=1}^{N} e_k^2)}$$

subject to:

$$\boxed{y_k[w^T\phi(x_k) + b] = 1 - e_k, \qquad k = 1, \ldots, N}$$

where $\phi(.)$: $\mathbb{R}^d \to \mathbb{R}^{d_h}$ is a nonlinear function which maps the d-dimensional input vector x from the input space to the d_h-dimensional feature space, possibly infinite.



The classifier in the dual space takes the form

$$\boxed{y(x) = \text{sign}[\sum_{k=1}^{N} \beta_k y_k K(x, x_k) + b]}$$

1 – 2-4 Performance Measures

For a binary classification model,

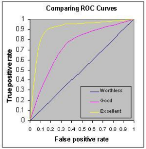
Confusion Matrix

	p' (predicted)	n' (predicted)
p (actual)	True Positive	True Negative
n (actual)	False Positive	False Negative

TP=True Positive
TN=True Negative
FP=False Positive
FN=False Negative

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



- The receiver operating characteristics (ROC) curve summarizes the performance of a classifier by showing the true positive rate (sensitivity) versus the false positive rate (1-specificity) as the discrimination threshold is varied.

2 – Outline

1. Chapter 1: Introduction
2. Chapter 2: Maximum Likelihood Estimation via Generalized Eigenvectors
3. Chapter 3: Robust PCA
4. Chapter 4: New Bandwidth Selection Criterion for KPCA
5. Chapter 5: Weighted LS-SVM Classifier
6. Chapter 6: Chemical Descriptor based on Connection-Table
7. Chapter 7:Conclusion and Future Studies

- 2-1 Generalized Eigenvalue Decomposition
- 2-2 Maximum Likelihood via Generalized Eigenvalue Decomposition
- 2-3 Case Study I
- 2-4 Case Study II
- 2-5 Conclusion

The work was published as Thomas, M., Daemen, A., De Moor, B., Maximum Likelihood Estimation of GEVD: Applications in Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume: 11, Issue: 4, 673 - 680: 2014

2 – 1 Generalized Eigenvalue Decomposition

- Let SVD of $\boxed{\text{A } (\text{B}^T B)^{-1/2}}$ be

$$\boxed{A(B^T B)^{-1/2} = PSQ^T}$$

Then

$$\boxed{(B^T B)^{-1/2} A^T A (B^T B)^{-1/2} Q = QS^T S}$$

where $S^T S = \Lambda$ and $W = Q$ with $Q^T Q = I_n$.

- Let define,

$$\boxed{D = A(B^T B)^{-1/2} Q = A(X^T)^{-1}}$$

with $\boxed{(\text{X})^{-1}(B^T B)(X^T)^{-1} = I_n}$ and columns of $(X^T)^{-1}$ are GEVs.

- **We use this relationship to identify maximum likelihood estimation via the generalized eigenvalue decomposition (MLGEVD).**

2 – 2 Maximum Likelihood Estimation via GEVD

We have,

$$\boxed{D = A(B^T B)^{-1/2} Q = A(X^T)^{-1}}$$

with $\boxed{(\mathrm{X})^{-1}(B^T B)(X^T)^{-1} = I_n}$ and columns of $(X^T)^{-1}$ are GEVs.

Referring to above Equation, MLGEVD problem which estimates the optimal generalized eigenvectors are formulated, as follows

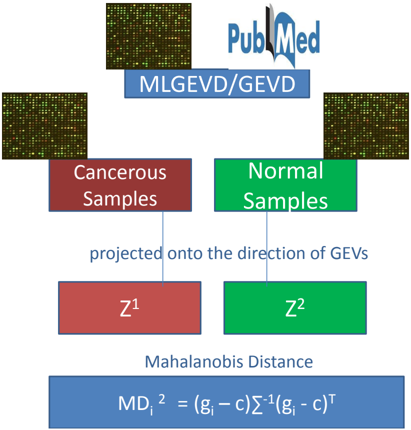
$$\text{minimize } ||D - A(X^T)^{-1}||^2$$
$$\text{subject to } [(X)^{-1}(B^T B)(X^T)^{-1}] = I_n$$

By solving Lagrangian of the above problem, we obtain the solution as

$$\tilde{r_i} = (A^T F_i^{-1} A + B^T B)^{-1} A^T F_i^{-1} D_i, i = 1 \ldots, n$$

It reveals an important mathematical property that in GSVD/GEVD framework, one of the data matrix acts as a prior information in the model development to obtain the generalized eigenvectors.

2 – 3 Case Study I: Identify Differentially Expressed Genes



MLGEVD/GEVD

Cancerous Samples

Normal Samples

projected onto the direction of GEVs

Z^1

Z^2

Mahalanobis Distance

$MD_i{}^2 = (g_i - c)\Sigma^{-1}(g_i - c)^T$

- Let $g_i = Z_i^1 - Z_i^2$ be the difference in score for gene i between normal and cancerous samples.
- The assumptions that the gene i with similar expression levels has approximately the same scores $Z_i^1 \approx Z_i^2$ and form a cloud of points around the origin.

2 – 3 Case Study I: Identify Differentially Expressed Genes

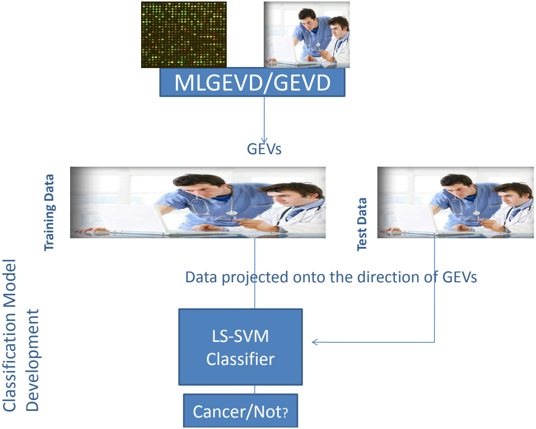
LS-SVM model: prediction of tumor and non-tumor samples for colon cancer

Genes selected by	kernel function	test AUC	p-value[a]
full data set	RBF	0.821(0.147)	0.019
GEVD	RBF	0.841(0.087)	0.072
MLGEVD	**RBF**	**0.895(0.060)**	
—-			

[a] *two-sided sign test for the comparison of full data sets and GEVD with MLGEVD.*

Results

- LS-SVM classifier on subsets of genes by MLGEVD offered the best prediction performances than LS-SVM classifier on whole gene sets and subsets of genes selected by GEVD framework.
- Out of 50 genes identified as differentially expressed by MLGEVD, majority of these genes are reported as top ranked genes in colon cancer in various studies.


MLGEVD/GEVD
GEVs
Training Data
Test Data
Classification Model Development
Data projected onto the direction of GEVs
LS-SVM Classifier
Cancer/Not?

2 – 4 Case Study II: Cancer Patient Classification

Predicting breast cancer based on clinical and microarray data.

	kernel function	linear	RBF	polynomial
Case I				
test AUC	MLGEVD	0.80(0.09)	0.79(0.01)	0.77(0.07)
	GEVD	0.77(0.08)	0.74(0.09)	0.63(0.02)
	p-value	0.03	0.01	2.72E-10
test F-score	MLGEVD	0.64(0.01)	0.57(0.02)	0.51 (0.13)
	GEVD	0.55(0.01)	0.49 (0.01)	0.26(0.03)
	p-value	0.17	0.23	0.03
Case II				
test AUC	MLGEVD	0.80(0.05)	0.75(0.09)	0.70(0.10)
	GEVD	0.79(0.06)	0.78(0.08)	0.61(0.06)
	p-value	0.01	0.04	0.01
test F-score	MLGEVD	0.60(0.03)	0.46 (0.06)	0.47(0.01)
	GEVD	0.56(0.02)	0.51(0.07)	0.50(0.05)
	p-value	0.02	0.06	0.11

p-value: two-sided sign test ; RBF: radial basis function

Results:

The LS-SVM classifier with the linear kernel function resulted in the best test AUC for both GEVD and MLGEVD.

With this kernel function, MLGEVD significantly outperformed GEVD for all three breast cancer case studies.

- This work shows the equivalence between MLGEVD and generalized ridge regression. This relationship reveals an important mathematical property of generalized eigenvalue decomposition (GEVD), in which the second argument acts as prior information in the model.
- We illustrate the importance of prior knowledge in clinical decision making/in identifying differentially expressed genes with case studies for which microarray data sets with corresponding clinical/literature information are available.
- The proposed approach could be used as an alternative of GEVD which significantly improved diagnosis, prognosis and prediction of therapy response.
- Both GEVD and MLGEVD used high-throughput data, which were difficult and expensive to collect only for model development.

3 – Outline

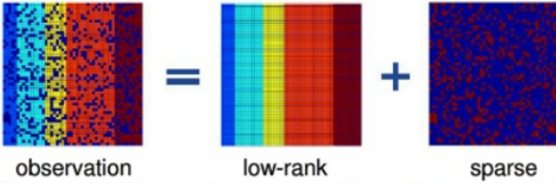
1. Chapter 1: Introduction
2. Chapter 2: Maximum Likelihood Estimation via Generalized Eigenvectors
3. Chapter 3: Robust PCA
4. Chapter 4: New Bandwidth Selection Criterion for KPCA
5. Chapter 5: Weighted LS-SVM Classifier
6. Chapter 6: Chemical Descriptor based on Connection-Table
7. Chapter 7:Conclusion and Future Studies

3 – Robust PCA to identify differentially expressed genes

- 3-1 Introduction
- 3-2 Robust PCA and Gene Expression Analysis
- 3-3 Conclusion

3 – 1 Introduction

The robust PCA proposed by Candes *et al* can recover a low-rank matrix C from highly corrupted measurements D.



Robust PCA solves the following optimization problem

$$\boxed{\min_{C,S} \|C\|_* + \lambda\|S\|_1,}$$

subject to $\boxed{\text{D=C+S}}$

where λ is a positive regulation parameter, $\|C\|_* = \Sigma_i \sigma_i(C)$ denote the nuclear norm of the matrix C, that is, the sum of its singular values, and $\|S\|_1 = \Sigma_{ij}|S_{ij}|$ denote the l_1-norm of S, which is efficient and robust to outliers.

Steps involved in:

- Identified differentially expressed genes using RPCA
- Search for co-expressed genes of differentially expressed ones using GeneMANIA, an online tool which finds other genes that are related to a set of input genes, using a very large set of functional association data.

LS-SVM classifier applied on whole data sets, subsets of genes obtained by robust PCA for predicting colon cancer.

Data Sources	No. of genes	Methods	Accuracy
Microarray and literature information	23	proposed approach	0.834(0.087)
Microarray	30	RPCA in Liu *et al.*	0.834(0.085)
Microarray and literature information	33	proposed approach and GeneMania	0.870(0.078)
Microarray	46	RPCA in *et al.* and GeneMania	0.795(0.095)
Microarray	2000	whole data set	0.834(0.097)

Results:

- The predictive performance has improved while considering both co-expressed genes and differentially expressed ones.
- The incorporation of literature information into microarray analysis improves the identification of disease associated genes.

- Robust PCA is a modification of the widely used statistical procedure of PCA which works well with respect to grossly corrupted observations.
- The proposed robust PCA approach on colon cancer data incorporating external knowledge into microarray analysis improves the identification of disease specific genes and performance of decision support in cancer diagnosis.
- The biological relevance and the prediction performance of selected genes emphasize the relevance of our approach for the identification of differentially expressed genes.
- This study show single data source is inadequate to explain complex network of genes underlying a disease.

4 – Outline

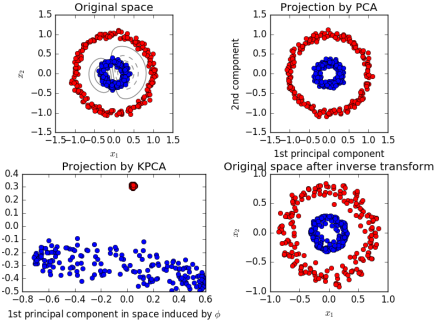
1. Chapter 1: Introduction
2. Chapter 2: Maximum Likelihood Estimation via Generalized Eigenvectors
3. Chapter 3: Robust PCA
4. Chapter 4: New Bandwidth Selection Criterion for KPCA
5. Chapter 5: Weighted LS-SVM Classifier
6. Chapter 6: Chemical Descriptor based on Connection-Table
7. Chapter 7:Conclusion and Future Studies

4 – New Bandwidth Selection Criterion Kernel PCA

- 4-1 Kernel PCA
- 4-2 Bandwidth selection criterion for kernel PCA
- 4-3 Results
- 4-4 Conclusion

The work was published as Thomas M., De Brabanter K., De Moor B.: New bandwidth selection criterion for Kernel PCA: Approach to Dimensionality Reduction and Classification Problems. BMC Bioinformatics 2014, 15:137 (2014).

4 – 1 Kernel PCA



Kernel PCA is defined as,

$$\boxed{\Omega_c \alpha = \lambda \alpha}$$

where $\Omega_{c,i,j}$ centered kernel matrix with RBF kernel function $K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2h^2})$ (RBF kernel with bandwidth h), is the eigenvector and λ is an eigenvalue.

$$\boxed{z_n(x) = \Sigma_{i=1}^{N} \alpha_i^{(n)} K(x_i, x)}$$

-$z_n(x)$ the score variable of sample x on n^{th} eigenvector $\alpha^{(n)}$.

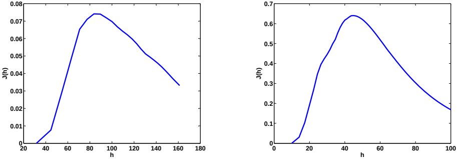
4 – 2 Bandwidth selection of Kernel PCA

Link with LS-SVM approach to kernel PCA and least squares cross validation in kernel density estimation, we proposed a data driver bandwidth selection criterion for kernel PCA. We propose the following tuning criterion for the bandwidth h:

$$\boxed{J(h) =_{h\in\mathbb{R}_0^+} \frac{1}{N}\sum_{j=1}^{N}\int |z_n^{(-j)}(x)|dx,}$$

where E denotes the expectation operator, N is the number of samples and $z_n^{(-j)}$ denotes the score variable with the jth observation is left out.

Figure: **Bandwidth Selection of KPCA for a Fixed Number of Components**



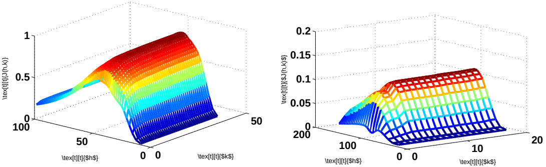
4 – 2 Bandwidth selection of Kernel PCA

The optimal bandwidth $h$ and number of components $k$ of kernel PCA can be obtained by

$$\boxed{J(h,k) =_{h\in\mathbb{R}_0^+, k\in\mathbb{N}_0} \frac{1}{N}\sum_{n=1}^{k}\sum_{j=1}^{N}\int |z_n^{(-j)}(x)|dx}$$

where N is the number of samples and $z_n^{(-j)}$ denotes the score variable with the jth observation is left out.

Figure: **Selection of optimal bandwidth and number of components for kernel PCA**



4 – 3 Results

Dataset	Kernel		preprocessing +	LS-SVM		PAM	Lasso
		whole data	PCA	KPCA	t-test(p<0.05)		
	RBF	0.769(0.127)	0.793(0.081)	0.822(0.088)	0.816(0.094)		
I	lin	0.822(0.068)	0.837(0.088)	0.864(0.078)	0.858(0.077)	0.787(0.097)	0.837(0.116)
	poly	0.818(0.071)	0.732(0.072)	0.825(0.125)	0.829(0.071)		
	RBF	0.637(0.146)	0.749(0.093)	0.780(0.076)	0.760(0.080)		
II	lin	0.803(0.059)	0.772(0.094)	0.790(0.075)	0.764(0.067)	0.659(0.084)	0.766(0.074)
	poly	0.701(086)	0.752(0.063)	0.753(0.072)	0.766(0.064)		
	RBF	0.832(0.143)	0.762(0.066)	0.879(0.058)	0.913(0.047)		
III	lin	0.915(0.043)	0.785(0.063)	0.878(0.066)	0.913(0.047)	0.707(0.067)	0.9359(0.0374)
	poly	0.775(0.080)	0.685(0.105)	0.8380(0.068)	0.913(0.047)		
	RBF	0.615(0.197)	0.853(0.112)	0.867(0.098)	0.853(0.187)		
IV	lin	0.953(0.070)	0.917(0.083)	0.929(0.077)	0.924(0.070)	0.759(0.152)	0.707(0.194)
	poly	0.762(0.118)	0.811(0.140)	0.840(0.131)	0.733(0.253)		

Results:

- As a preprocessing step kernel PCA outperformed PCA in the performance of LS-SVM classifier on all case studies.
- The feature selection techniques such as t-test, Lasso and PAM preformed well on few cases, but the number of genes selected on each iterations widely varied.
- Considering the possibility of increasing size and complexity of microarray data sets in future, dimensionality reduction and nonlinear techniques have its own significance.

4 – 4 Conclusion

- We propose a data-driven bandwidth selection criterion for kernel PCA by tuning the optimum bandwidth and the number of principal components.
- The performance of the proposed strategy is significantly better than an existing optimization algorithm for kernel PCA
- Its classification performance is not sensitive to any number of selected genes, so the proposed method is more stable than others proposed in literature
- It reduces the dimensionality of the data while keeping as much information as possible of the original data

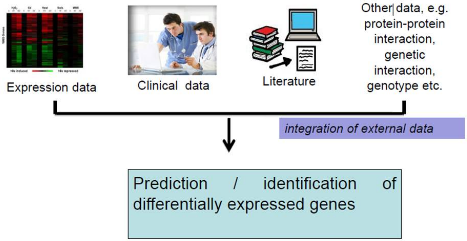
5 – Outline

1. Chapter 1: Introduction
2. Chapter 2: Maximum Likelihood Estimation via Generalized Eigenvectors
3. Chapter 3: Robust PCA
4. Chapter 4: New Bandwidth Selection Criterion for KPCA
5. **Chapter 5: Weighted LS-SVM Classifier**
6. Chapter 6: Chemical Descriptor based on Connection-Table
7. Chapter 7:Conclusion and Future Studies

- 1-1 Introduction
- 1-2 Kernel GEVD
- 1-3 Weighted LS-SVM classifier
- 1-4 Predicting breast cancer using an expression values weighted clinical classifier
- 1-5 Results
- 1-6 Conclusion

This paper was published in BMC Bioinformatics: Thomas M., De Brabanter K., Suykens J.A.K., De Moor B.: Predicting breast cancer using an expression values weighted clinical classifier. BMC Bioinformatics 2014, 15:6603 (2014).

5 – 1 Introduction



Expression data

Clinical data

Literature

Other data, e.g. protein-protein interaction, genetic interaction, genotype etc.

integration of external data

Prediction / identification of differentially expressed genes

- 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.
- 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.
- LS-SVM is a powerful classifier in microarray analsyis
- 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.

- 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} column of m× n matrix A and p× n matrix B respectively).
- Consider the feature maps $\varphi^{(1)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_1}$ and $\varphi^{(2)}(.) : \mathbb{R}^p \to \mathbb{R}^{n_2}$ 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

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

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

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

The kernel GEVD is described as follows:

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

where $\lambda = \frac{1}{\gamma}$ eigenvalue, $\Omega_c^{(1)}$, $\Omega_c^{(2)}$ are centered kernel matrices and α are generalized eigenvectors.

5 – 3 Weighted LS-SVM Classifier

Given the link between LS-SVM approach to kernel GEVD and the weighted LS-SVM classifier, we proposed a mathematical framework which representing the kernel GEVD in the form of weighted LS-SVM classifier.

In primal space the problem:

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

$$\left[\begin{array}{cc} 0 & 1_N^T \\ 1_N & \Omega_c^{(1)} + \gamma^{-1} \Omega_c^{(2)} \end{array}\right] \left[\begin{array}{c} b \\ \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ y \end{array}\right]$$

In dual space,

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

with α_i are the Lagrange multipliers, γ is a regularization parameter chosen by the 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.

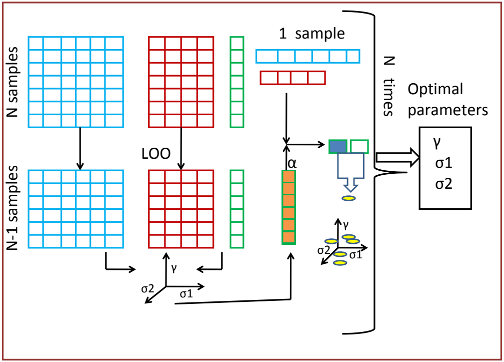


Figure: Optimization algorithm for the weighted LS-SVM classifier. The data sets represented as matrices with rows corresponding to patients and columns corresponding to genes and clinical parameters respectively for first and second data sets. v-fold (in this figure for simplicity we assumes $v = 1$) cross validation is applied to select the optimal parameters for the LS-SVM classifier.

Comparisons of Different Classifiers, on Single Data source vs. Multiple Data Sources

Classifiers	Case I	Case II	Case III	Case IV
CL +LS-SVM				
test AUC	0.7795(0.0687)	0.7979(0.1358)	0.6152(0.0565)	0.6622(0.0628)
p-value	0.0511	0.0609	0.0086	2.2790E-09
MA+LS-SVM				
test AUC	0.7001(0.0559)	0.8060(0.0728)	0.6216(0.0348)	0.7357(0.0085)
p-value	0.0001	0.0010	00.0040	3.4954E-05
GEVD+LS-SVM				
test AUC	0.7716(0.0818)	0.7685(0.0645)	0.6172(0.0426)	0.7528(0.1257)
p-value	0.0756	2.0196E-05	0.0040	0.0722
KGEVD+LS-SVM				
test AUC	0.7936(0.0650)	0.821(0.0670)	0.6539(0.0616)	0.773(0.1011)
p-value	0.1701	0.0525	0.0064	0.1485
weighted LS-SVM				
test AUC	0.8177(0.0666)	0.8465(0.0480)	0.6921(0.0327)	0.8119(0.0893)

p-value: One-sided paired-sampled t-test for the comparison of weighted LS-SVM with other classifiers.
CL and MA are the clinical and microarray kernels of RBF kernel functions.

Results:

- On all case studies, weighted LS-SVM classifier outperformed all other discussed methods (LS-SVM on RBF clinical kernel and microarray kernel individually, GEVD and kernel GEVD as pre-processing step,followed by LS- SVM on reduced data), in terms of test AUC.
- The weighted LS-SVM performance is slightly better on the second and fourth cases, but not significantly, than the kernel GEVD.

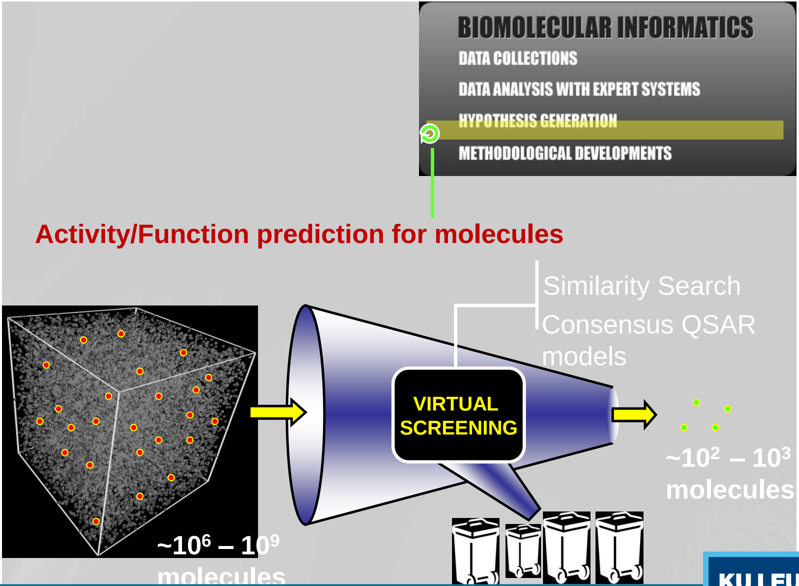
- The ultimate goal of this work is to propose a machine learning approach which is functional in both data fusion and supervised learning.
- We further analyzed the potential benefits of merging microarray and clinical data sets for prognostic application in breast cancer diagnosis.
- A clinical classifier weighted with microarray data set results in significantly improved diagnosis, prognosis and prediction responses to therapy.
- The proposed model has been shown to be a promising mathematical framework in both data fusion and non-linear classification problems.
- Possible additional applications of weighted LS-SVM include integration of genomic information collected from different sources and biological processes.

1. Chapter 1: Introduction
2. Chapter 2: Maximum Likelihood Estimation via Generalized Eigenvectors
3. Chapter 3: Robust PCA
4. Chapter 4: New Bandwidth Selection Criterion for KPCA
5. Chapter 5: Weighted LS-SVM Classifier
6. Chapter 6: Chemical Descriptor based on Connection-Table
7. Chapter 7:Conclusion and Future Studies

6 – Chemical Descriptor based on Connection-Table

- 1-1 Motivation
- 1-2 Molecular Descriptors
- 1-3 PCA conversion of the connection-table and weighted LS-SVM
- 1-4 Weighted LS-SVM classifier
- 1-5 Case Study
- 1-6 Results
- 1-7 Conclusion

6 – 1 Motivation



BIOMOLECULAR INFORMATICS

- DATA COLLECTIONS
- DATA ANALYSIS WITH EXPERT SYSTEMS
- HYPOTHESIS GENERATION
- METHODOLOGICAL DEVELOPMENTS

Activity/Function prediction for molecules

Similarity Search

Consensus QSAR models

VIRTUAL SCREENING

~10^6 – 10^9 molecules

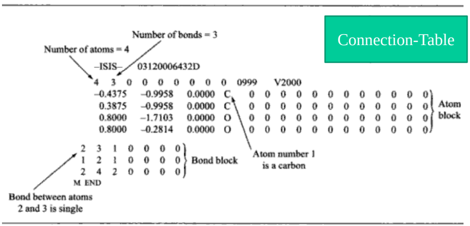
~10^2 – 10^3 molecules

Molecular descriptors are numerical values that characterize properties of molecules. The descriptors fall into Four classes

- Topological
- Geometrical
- Electronic
- Hybrid or 3D Descriptors

Connection Table a portion of a structure-data File (SDF File):

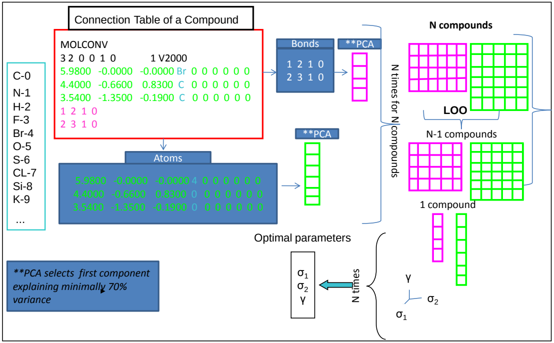
A chemical file format: MOL file



```
  –ISIS–   03120006432D
  4  3  0  0  0  0  0  0  0  0999   V2000
 –0.4375  –0.9958   0.0000  C   0  0  0  0  0  0  0  0  0  0  0  0
  0.3875  –0.9958   0.0000  C   0  0  0  0  0  0  0  0  0  0  0  0
  0.8000  –1.7103   0.0000  O   0  0  0  0  0  0  0  0  0  0  0  0
  0.8000  –0.2814   0.0000  O   0  0  0  0  0  0  0  0  0  0  0  0
2  3  1  0  0  0  0
1  2  1  0  0  0  0
2  4  2  0  0  0  0
M  END
```

Connection-Table
Number of bonds = 3
Number of atoms = 4
Atom block
Bond block
Atom number 1 is a carbon
Bond between atoms 2 and 3 is single

An overview of chemical descriptor formation from the connection table of compounds


Connection Table of a Compound

```
MOLCONV
3 2  0  0  1  0          1 V2000
5.9800  -0.0000  -0.0000 Br  0  0  0  0  0  0
4.4000  -0.6600   0.8300  C   0  0  0  0  0  0
3.5400  -1.3500  -0.1900  C   0  0  0  0  0  0
1  2  1  0
2  3  1  0
```

**PCA selects first component explaining minimally 70% variance

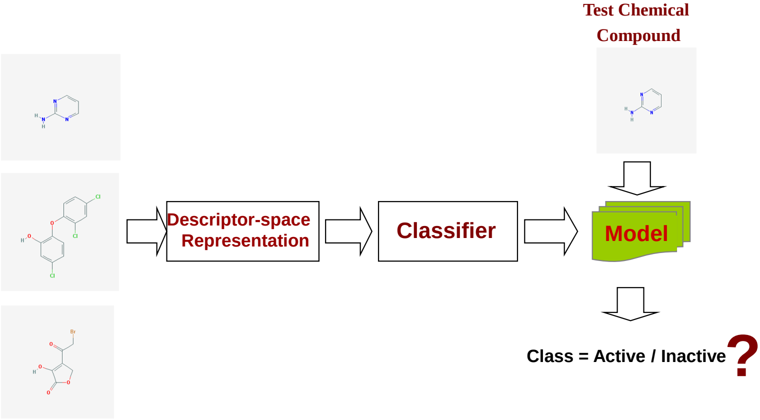
- PCA is applied to the connection-table of each compounds to define a new structural descriptor in terms of two vectors. This results into two matrices: atoms vs. compounds and bonds vs. compounds.
- The weighted LS-SVM framework integrates these two vectors into a single vector named as weighted chemical descriptor and performs further prediction.
- LOO-CV is applied to select the optimal parameters.

Weighted LS-SVM classifier, a mathematical framework for integrating two data sources are defined as ,

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

with α_i are the Lagrange multipliers, γ is a regularization parameter chosen by the 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.

Identify new compounds which inhibit biofilms formed by either Salmonella or Pseudomonas.



Test Chemical Compound

Descriptor-space Representation

Classifier

Model

Class = Active / Inactive?

6 – 6 Result

Comparison of averaged classification performances of different descriptors: proposed descriptor, MACCS keys, ECF, Path keys and BCUT descriptors to identify the active and inactive compounds in *Salmonella* and *Pseudomonas* biofilm

	Proposed Method	MACCS Keys	ECF	Path keys	BCUT
Salmonella					
Accuracy(std)	0.8071(0.0581)	0.7706(0.0500)	0.7686(0.0438)	0.7735(0.0454)	0.766(0.04)
p-value		0.0100	0.0100	0.0005	0.012
Test AUC(std)	0.6453(0.0409)	0.6988(0.0294)	0.6525(0.0263)	0.6425(0.0328)	0.665 (0.077)
p-value		6.73E-05	0.3528	0.2800	0.031
F-score(std)	0.3058(0.0375)	0.0212(0.0333)	0.0315(0.0241)	0.036(0.0142)	0.484(0.0823)
p-value		1.29E-16	5.52E-17	1.00E-17	0.064
Pseudomonas					
Accuracy(std)	0.8179(0.0492)	0.7908(0.0441)	0.7829(0.0360)	0.7882(0.0313)	0.65(0.055)
p-value		0.0001	0.0019	0.0004	0.012
Test AUC(std)	0.6549(0.0267)	0.7041(0.0258)	0.6559(0.0397)	0.7118(0.0395))	0.588(0.066)
p-value		5.06E-06	0.9104	3.30E-05	0.212
F-score(std)	0.3277(0.0346)	0.0211(0.0330)	0.1401(0.0854)	0.0874(0.0469)	0.743(0.059)
p-value		2.34E-17	1.38E-07	8.71E-14	0.024

Results:

- In both case studies, the proposed weighted LS-SVM based descriptor performed well in terms of test accuracy and F-score.
- While the best test AUC returned by MACCS keys for Salmonella and Path keys for Ps

- PCA was used to decompose the connection-table of each compound effectively to a low dimensional representation, as such defining a new structural descriptor of chemical compounds.
- A weighted LS-SVM approach was used to design a weighted chemical descriptor and to predict the biological activity of chemical compounds.
- The results illustrate that the obtained descriptor offers an improved model to identify very active compounds in a specific biological condition.
- The newly proposed approach, the weighted chemical descriptors of molecular structure, identified accurately the inhibitors on Salmonella, Pseudomonas biofilms formation, Thrombin, Trypsin and FactorXa, than other discussed descriptors.
- The proposed machine learning technique could be applicable to any classification/prediction problem which is based on the molecular structure of compounds.

1. Chapter 1: Introduction
2. Chapter 2: Maximum Likelihood Estimation via Generalized Eigenvectors
3. Chapter 3: Robust PCA
4. Chapter 4: New Bandwidth Selection Criterion for KPCA
5. Chapter 5: Weighted LS-SVM Classifier
6. Chapter 6: Chemical Descriptor based on Connection-Table
7. Chapter 7:Conclusion and Future Studies

- Microarray data, which was difficult and expensive to collect were incorporated as prior information into clinical decision-making, improving the classification performance and offering better diagnosis and prognosis.
- Incorporation of literature information into microarray analysis improved the possibility for obtaining stable disease associated genes.
- The linear projections based on GEVD will not perform very well in the model development, if the primary data source contains only limited number of features.
- To tackle these problems either we have to perform the projection based on kernel based GEVD or remove irrelevant features from the prior data using feature selection techniques.
- The unsupervised dimensionality reduction methods are most useful in the practical applications in which the labeled data are usually expensive to collect.
- We offers a data driven bandwidth selection criterion for KPCA which is executing in an unsupervised mode.
- We proposed a kernel-based mathematical framework for data integration and classification: a weighted LS-SVM classifier.
- This approach could be considered as a standard mathematical problem to produce better classification performance based on heterogeneous data integration.

7 – Future Studies

- The big data technologies process large quantities of data within tolerable elapsed time. In future one can propose the implementation of GEVD and MLGEVD for these platform and overcome the technical challenges with big data.
- If the data set grows or changes over time, the RPCA algorithm needs to run from scratch. This raises the question of how the existing models can be extended to include the scalable data sets.
- Advanced analytics techniques are needs to be formulated for multivariate statistics, such as kernel PCA and kernel regression, in matrix form over big data platforms.
- The chemoinformatics work will be extended to address the questions such as the parameters which are important for activity and how to modify molecules to improve the activity in a specific biological condition etc.

Thank
YOU!