Kernel-based data fusion for machine learning: methods and applications in bioinformatics and text mining

Jury:
Prof. dr. ir. H. Hens, chairman
Prof. dr. ir. B. De Moor, promoter
Prof. dr. ir. Y. Moreau, co-promoter
Prof. dr. ir. G. Bontempi
Prof. dr. ir. T. De Bie
Prof. dr. L. Dehaspe
Prof. dr. ir. P. Dupont
Prof. dr. ir. J. Suykens

Shi Yu

24-11-2009
Thermotechnical Institute
Kasteelpark Arenberg 41
K.U.Leuven

Overview

- General background
- Main topics (main contributions)
  - Kernel fusion for one class problem
  - Kernel fusion for multi-class problem
  - Kernel fusion for large scale data
  - Kernel fusion for clustering analysis
- Conclusions and future research

Overview

- General background
- Main topics (main contributions)
- Conclusions

History of learning

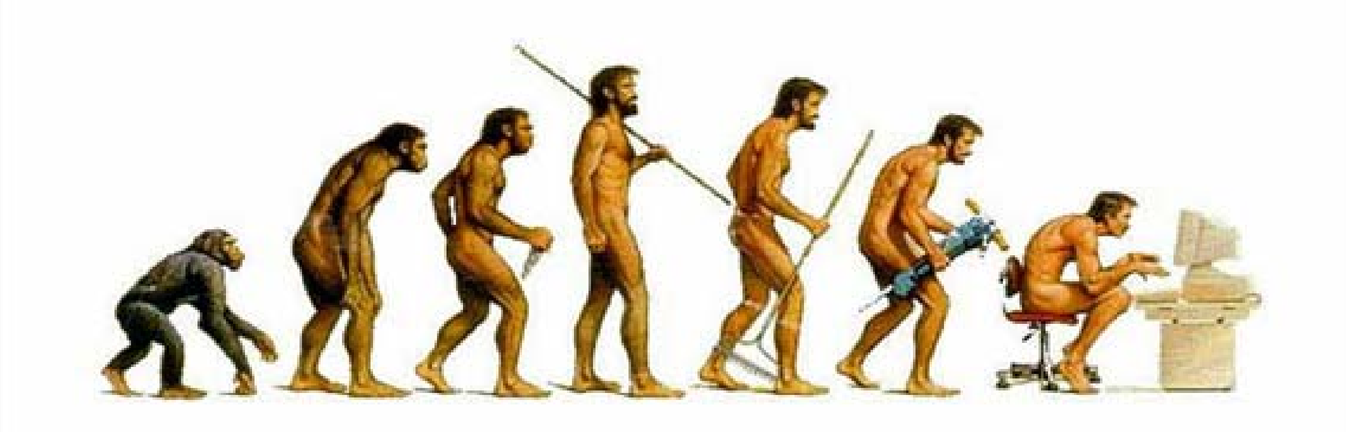


Image from http://www.buildamovement.com/blog/wp-content/uploads/2009/09/evolution.jpg

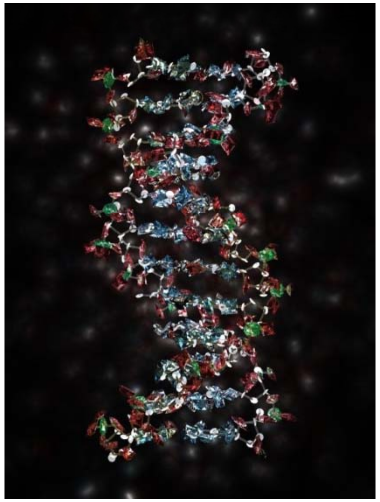
- Learning in the jungle …… Learning through machines …
- The connections between machine learning and biology
  - evolutionary computing, perceptron, neural networks, …
- Big breakthroughs in the community = tiny steps in the history

An expedition towards the real computational intelligence

ISS Expedition 9, NASA, adapted from
apod.nasa.gov/apod/image/0408/supply_iss.jpg

Robotics

Figure adapted from
http://www.gadgets-reviews.com



Synthetic Biology

Figure adapted from
http://adamant.typepad.com

Pattern Analysis

Figure adapted from
http://treepax.com/blog/2007/11/20/girls-face/

Adaptability and exquisiteness of biological intelligence (1)

- Learning from multiple senses
- Integration and prioritization of the senses

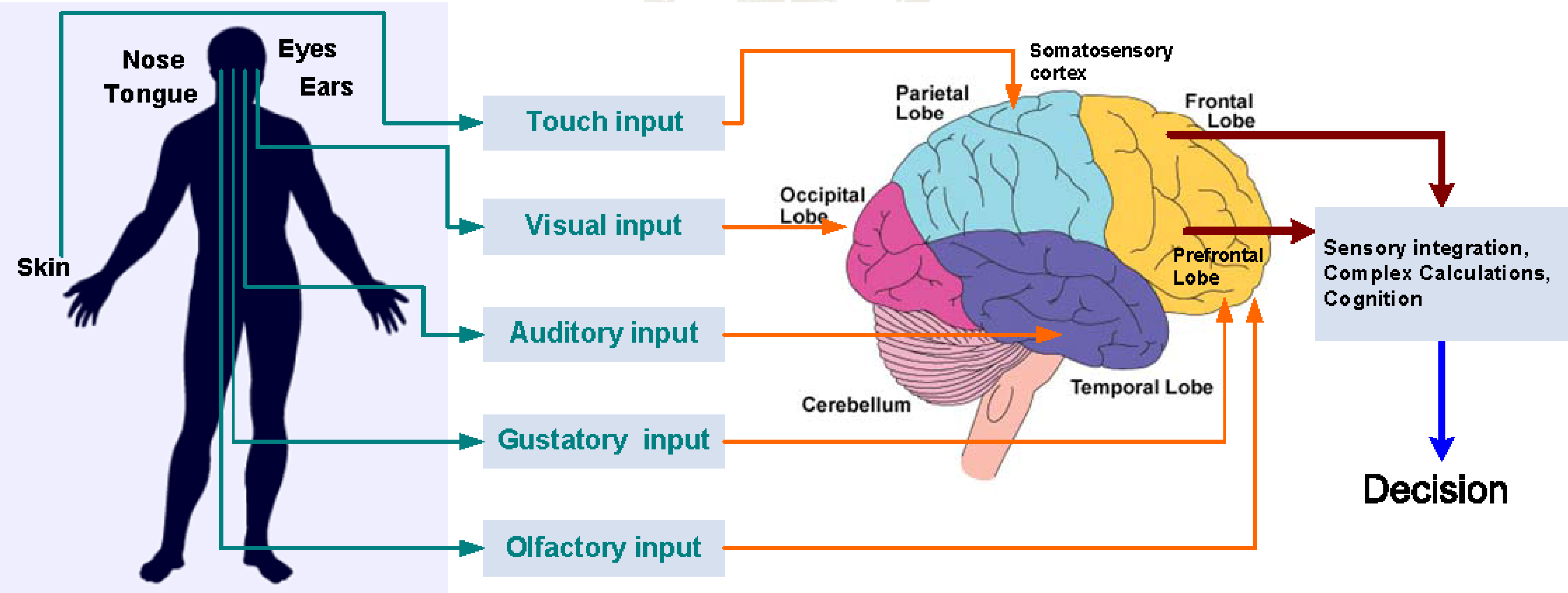
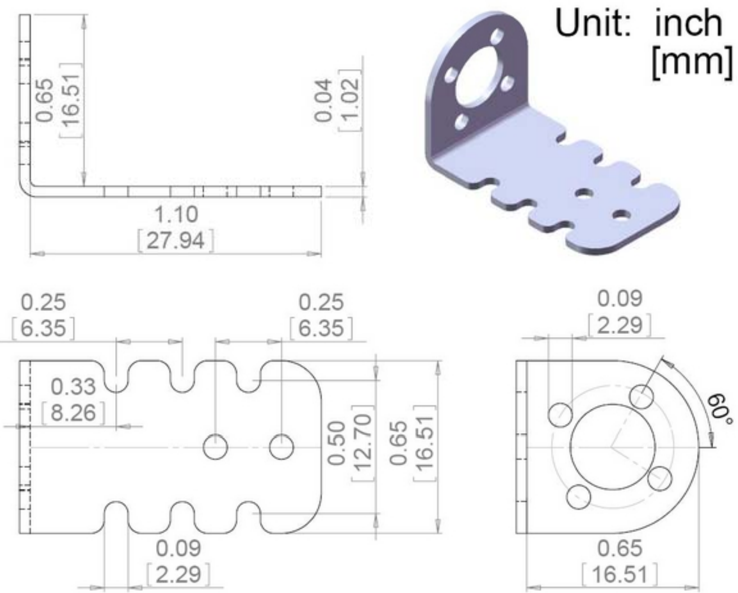


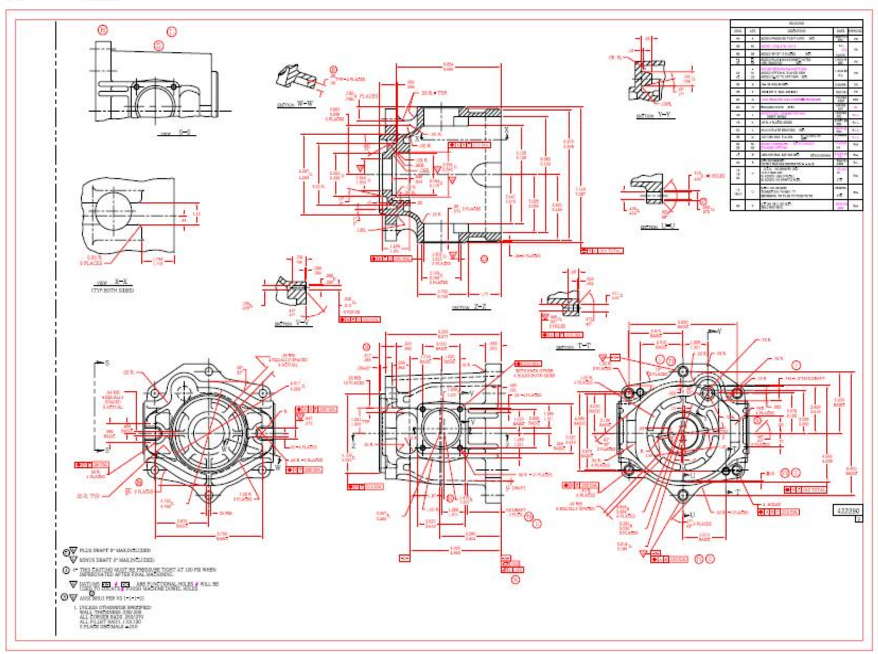
Figure of human body is adapted from Widen Clinic, http://www.widenclinic.com/health.html
Figure of brain is adapted from Jodi house, www.cycrout.com/jillianpwinslow/brain.jpg

Adaptability and exquisiteness of biological intelligence (2)

- Geometry reconstruction in a higher dimensional space using multiple lower dimensional projections
- Multi-view learning



Pololu Mini Metal Gearmotor Bracket Pair
Figure adapted from www.pololu.com



Mechanical Sample Drawing
Figure adapted from sofeon.com/mechanical.jpg

From biological intelligence to machine intelligence

- Human brains: abstract knowledge, high-level inference
- Machine learning and data mining: statistical and numerical patterns
- Incorporating more information sources in machine learning and data mining may be beneficial
  - to improve the statistical significance
  - to leverage the interactions and correlations between multiple observations
  - to reduce the noise
  - to obtain more refined and high-level knowledge

Multi-source machine learning: historical background

- Canonical correlation (Hotelling, 1936)
- Inductive logic programming (Muggleton and De Raedt, 1994) and multi-source learning search space (Fromont *et al.*, 2005)
- Additive model and ensemble learning
  - Bagging (Breiman, 1996) and Boosting (Freund and Schapire, 1997)
  - Feed-forward neural network ensemble (Drucker *et al.*, 1993)
- Bayesian networks integration (Pearl, 1988; Gevaert, 2008)
- Kernel-based data fusion (Vapnik, 1998; Boser *et al.*, 1992; Bach *et al.*, 2003; Lanckriet *et al.*, 2004)

Kernel methods primer

- Dual representation, kernel trick and feature (Hilbert) space
  - Bayesian network learning: training data → model parameters
  - Kernel methods: keeping the training data during the prediction phase as a metric defined as the inner product (*dual representation*) of any two data points
  - Linear input space $\mathbb{R}$ → Nonlinear embeddings in a higher dimensional feature space (Hilbert space) $\mathcal{F}$
  - The inner product of the embedded data is specified via a kernel function $K(\vec{x}_1, \vec{x}_2) = \phi(\vec{x}_1)^T \phi(\vec{x}_2)$, known as the *kernel trick*

Kernel methods for classification

- Support Vector Machines

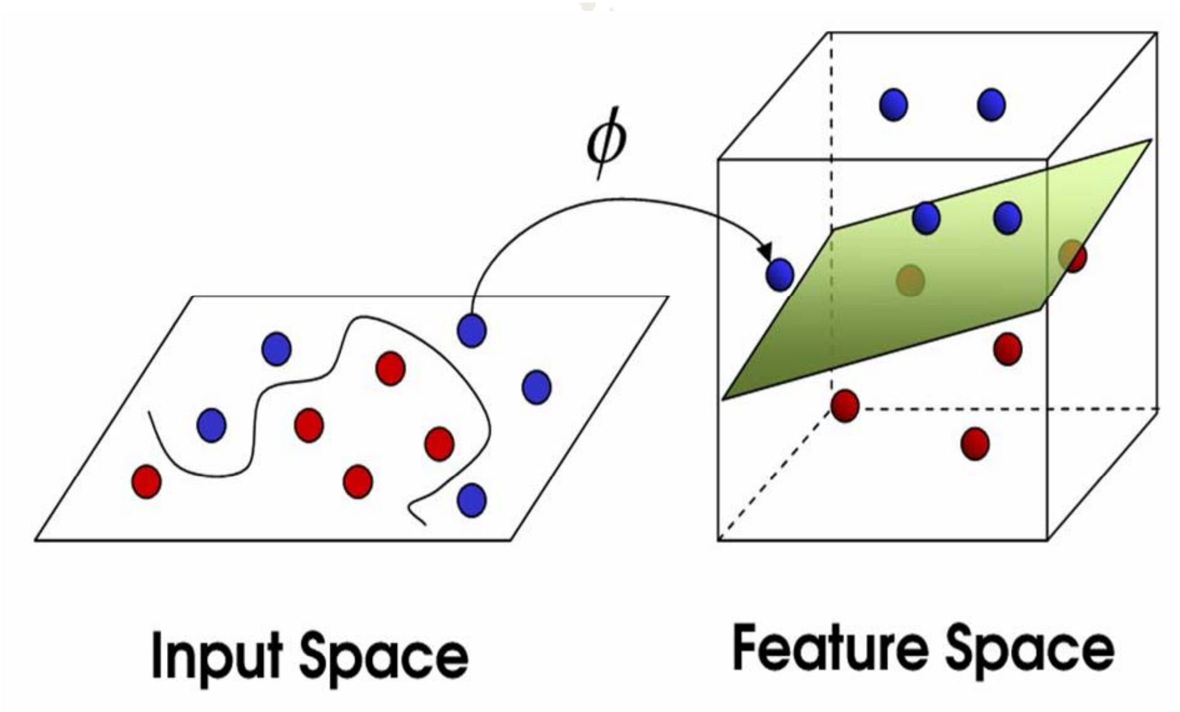
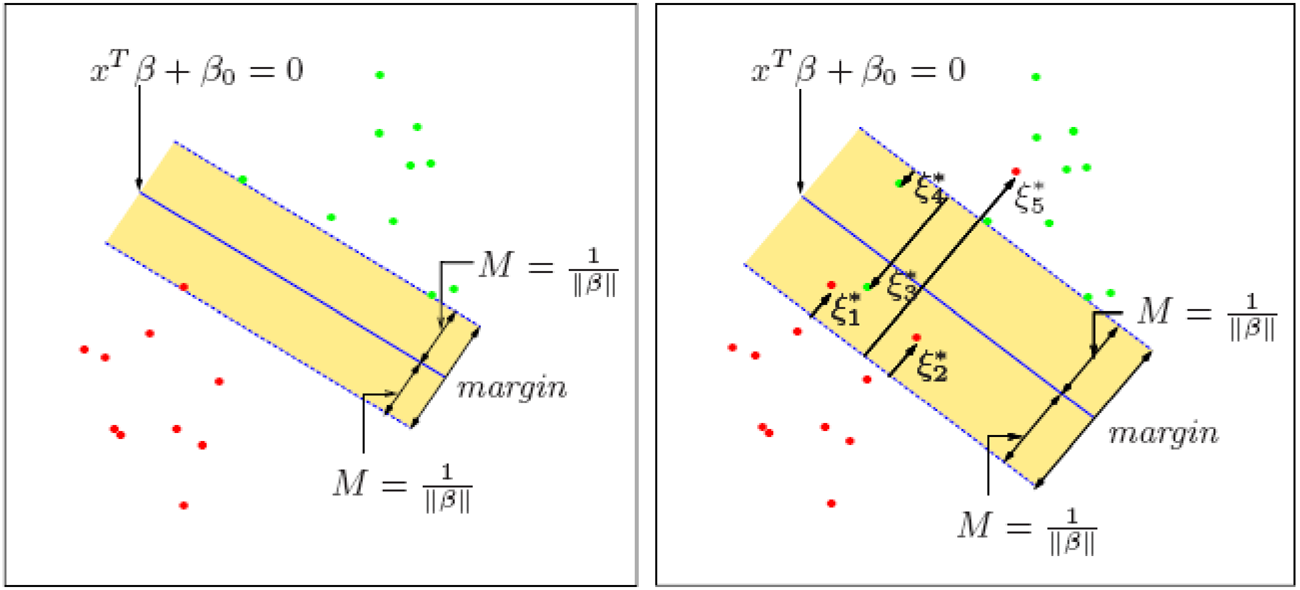


Figure adapted from www.imtech.res.in/raghava/rbpred/svm.jpg

Kernel methods for classification



Vapnik's SVM

Figure adapted from Hastie *et al.*, 2009

$$\boxed{\text{P:}} \min_{\vec{w},b,\vec{\xi}} \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{k=1}^{N}\xi_k$$

$$\text{s.t.} \quad y_k[\vec{w}^T\phi(\vec{x}_k) + b] \geq 1 - \xi_k, \quad k = 1, ..., N$$

$$\xi_k \geq 0, \quad k = 1, ..., N.$$

$$\boxed{\text{D:}} \max_{\vec{\alpha}} \; -\frac{1}{2}\sum_{k,l=1}^{N}\alpha_k\alpha_l y_k y_l \phi(\vec{x}_k)^T\phi(\vec{x}_l) + \sum_{k=1}^{N}\alpha_k$$

$$\text{s.t.} \quad 0 \leq \alpha_k \leq C, \quad k = 1, ..., N$$

$$\sum_{k=1}^{N}\alpha_k y_k = 0.$$

Kernel methods for data fusion

- Additive expansion of the prediction function

$$y_k\left[\sum_{j=1}^{p}\left(\sqrt{\theta_j}\vec{w}_j^T\phi_j(\vec{x}_k)\right)+b\right]\geq 1-\xi_k,\quad k=1,...,N$$

- Denote $\vec{\eta}_j=\sqrt{\theta_j}\vec{w}_j$, we have:

$$\boxed{\text{P:}}\quad \min_{\vec{\eta},b,\vec{\theta},\vec{\xi}}\ \frac{1}{2}\sum_{j=1}^{p}\vec{\eta}_j^T\vec{\eta}_j+C\sum_{k=1}^{N}\xi_k$$

$$\text{s.t.}\quad y_k\left[\sum_{j=1}^{p}\left(\vec{\eta}_j^T\phi_j(\vec{x}_k)\right)+b\right]\geq 1-\xi_k,\quad k=1,...,N$$

$$\xi_k\geq 0,\ \sum_{k=1}^{N}\xi_k=C,\quad k=1,...,N$$

$$\boxed{\text{D:}}\quad \min_{\vec{\theta}}\max_{\vec{\alpha}}\ -\frac{1}{2}\sum_{k,l=1}^{N}\alpha_k\alpha_l y_k y_l\boxed{\sum_{j=1}^{p}\left(\theta_j K_j(\vec{x}_k,\vec{x}_l)\right)}+\sum_{k=1}^{N}\alpha_k$$

$$\text{s.t.}\quad 0\geq\alpha_k\geq C,\quad k=1,...,N$$

$$\sum_{k=1}^{N}\alpha_k y_k=0,$$

$$\theta_j\geq 0,\ \sum_{j=1}^{p}\theta_j=1,\quad j=1,...,p,$$

- In a dual representation, the additive expansion of SVMs on multiple data source is denoted as *kernel fusion*

Loss functions in kernel methods

- In SVMs, there are many criteria to assess the quality of predictions based on observations

$$\min_{\vec{w}} \frac{1}{2}\vec{w}^T\vec{w} + \lambda \sum_{k=1}^{N} L[y_k, f(\vec{x}_k)],$$

Loss Function	$L[y, f(\vec{x})]$	Classifier name
Binomial Deviance	$log[1 + e^{-yf(\vec{x})}]$	logistic regression
Hinge Loss	$\lvert 1 - yf(\vec{x}) \rvert_+$	SVM
Squared Error	$[1 - yf(\vec{x})]^2$ (equality constraints)	LS-SVM
L_2 norm	$[1 - yf(\vec{x})]^2$ (inequality constraints)	2-norm SVM
Huber's Loss	$\begin{cases} -4yf(\vec{x}), & yf(\vec{x}) < -1 \\ [1 - yf(x)]^2, & \text{otherwise} \end{cases}$	

Kernel methods for clustering

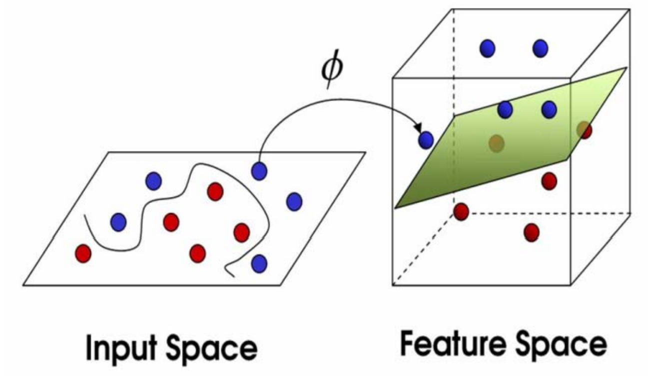
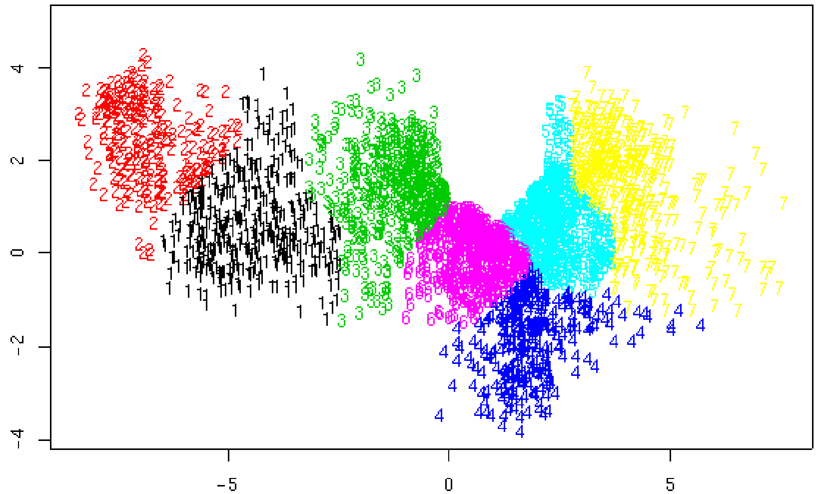


Figure adapted from www.imtech.res.in/raghava/rbpred/svm.jpg



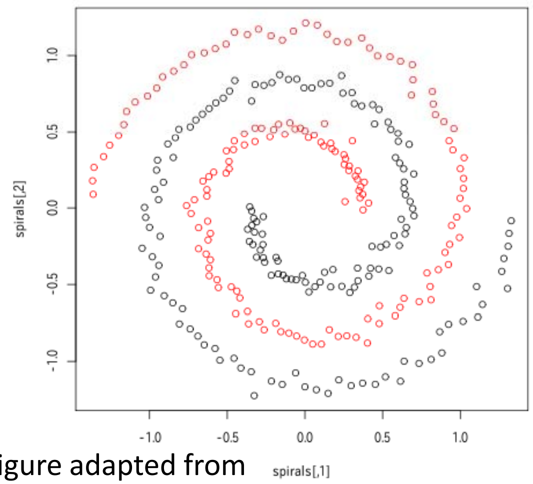


Figure adapted from http://bm2.genes.nig.ac.jp/RGM2/R_current/library/kernlab/man/images/specc_001.png

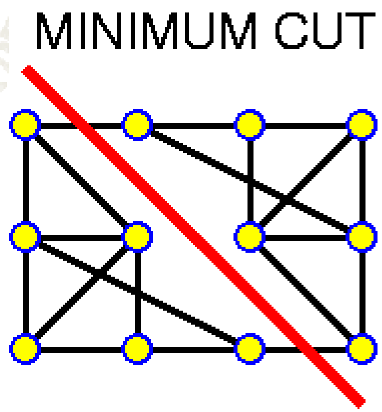
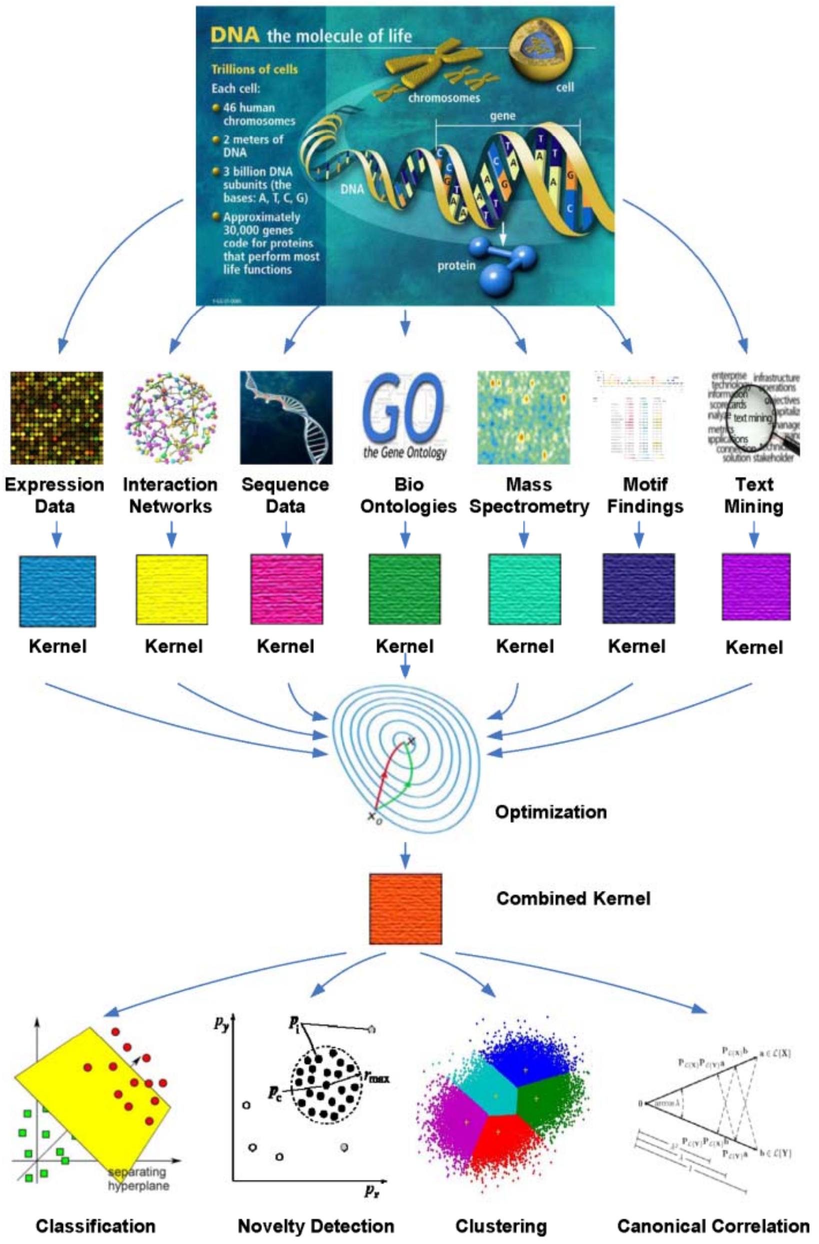


Figure adapted from http://users.informatik.uni-halle.de/~jopsi/drand04/mincut.gif

A bioinformatics perspective



DNA image from the U. S. Department of Energy Human Genome Project
Microarray image from bio.davidson.edu
Interaction network image from systemsbiology.org.au
Sequence image from firstscience.com
GO image from geneontology.org
Motif image from http://www.pnas.org/content/102/33/11651/F1.medium.gif
Text mining image from arkabio.com
Optimization image from commons.wikimedia.org
Classification image adapted from Van Looy *et al. Critical Care* 2007 **11**:R83 doi:10.1186/cc6081
Clustering image adapted from http://www.mathworks.com/matlabcentral/fx_files/19344/1/preview.jpg

Rayleigh quotient problems in machine learning algorithms

- The (general) Rayleigh quotient (RQ) type problem

$$\max_{\vec{w}} \frac{\vec{w}^T A \vec{w}}{\vec{w}^T B \vec{w}}, \text{ or } \min_{\vec{w}} \frac{\vec{w}^T A \vec{w}}{\vec{w}^T B \vec{w}}, \quad A \succeq 0, B \succ 0$$

- Principal Component Analysis, Canonical Correlation Analysis, Fisher Discriminant Analysis, K-means clustering, spectral clustering, Kernel Laplacian clustering, one class SVM, least squares SVM, Partial least squares …
- The solution to the RQ type problem can be straightforwardly extended to a set of algorithms

Kernel fusion for RQ-type problems

- Problem statement

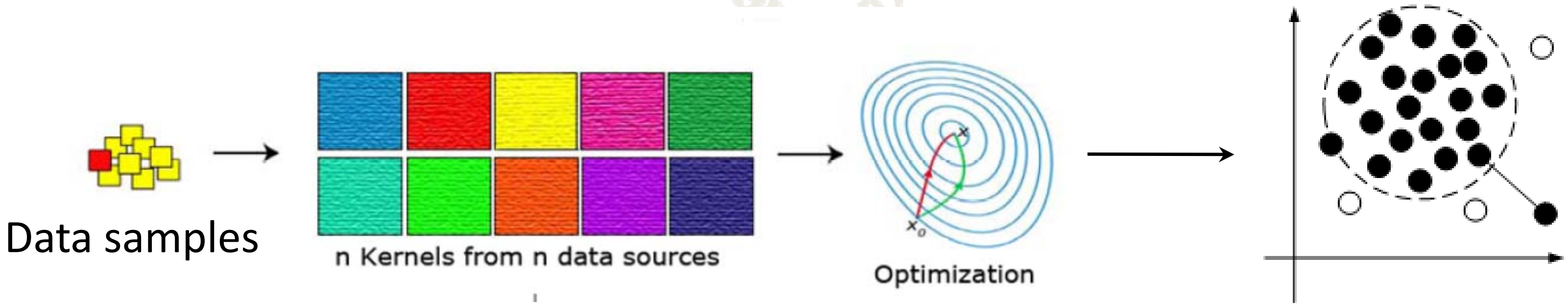
$$\max_{\vec{w}} \quad \frac{\vec{w}^T \Omega \vec{w}}{\vec{w}^T \vec{w}},$$
$$\text{s.t.} \quad \Omega = \left\{ \sum_{j=1}^{p} \theta_j K_j | \sum_{j=1}^{p} \theta_j^\delta = 1, \forall j, \theta_j >= 0, \delta > 0 \right\},$$
$$\vec{w}^T \vec{w} = 1.$$

- Sparse($\delta = 1, L_\infty$-norm) and non-sparse ($\delta = 2, L_2$-norm) solutions for data fusion
- Automatic optimization of θ_j

Overview

- General background
- Main topics (main contributions)
- Conclusions

Topic 1: Kernel fusion for one class problem: algorithms and applications



Fundamental problem

$$\max_{\theta} \min_{\vec{\alpha}} \ \vec{\alpha}^T \left(\sum_{j=1}^{p} \theta_j Q_j \right) \vec{\alpha}$$
$$\text{s.t.} \ \ Q_j \succeq 0, \quad j = 1, ..., p$$
$$\theta_j \geq 0, \quad j = 1, ..., p$$
$$\sum_{j=1}^{p} \theta_j^{\delta} = 1.$$

$\boxed{\delta = 1}$

$$\min_{\vec{\alpha}, t} \ \ t$$
$$\text{s.t.} \ \ Q_j \succeq 0, \quad j = 1, ..., p$$
$$t \geq \vec{\alpha}^T Q_j \vec{\alpha}, \quad j = 1, ..., p.$$

$$L_\infty : \ t^* = ||\vec{\alpha}^T Q_j \vec{\alpha}||_\infty = \max\{\alpha^T Q_1 \vec{\alpha}, ..., \alpha^T Q_p \vec{\alpha}\}.$$

$\boxed{\delta = 2}$

$$\min_{\vec{\alpha}, \eta} \ \ \eta$$
$$\text{s.t.} \ \ Q_j \succeq 0, \quad j = 1, ..., p$$
$$s_j \geq \vec{\alpha}^T Q_j \vec{\alpha}, \quad j = 1, ..., p$$
$$\eta \geq ||s_j||_2, \quad j = 1, ..., p.$$

$$L_2 : \quad \eta^* = ||\vec{\alpha}^T Q_j \vec{\alpha}||_2 \ .$$

Kernel coefficients in kernel fusion

- Let's denote the combined matrix as

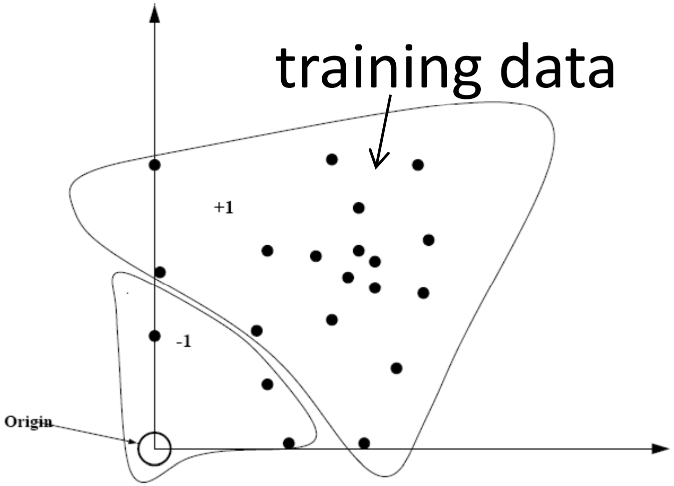
$$\Omega = \left\{ \sum_{j=1}^{p} \theta_j K_j \middle| \forall j,\ \theta_j \geq 0,\ K_j \succeq 0,\ \sum_{j=1}^{p} \theta_j^{\delta} = 1 \right\} .$$

Table 1: The notations used in the thesis are based on the dual problem and they are linked to the equivalent notations in the primal problem

	primal problem	dual problem
variable	θ_j	$\vec{\alpha}^T K_j \vec{\alpha}$
$\mathbf{L}_\infty$	$\lvert\theta_j\rvert = 1\ (\delta = 1)$	$\max \lvert\lvert\vec{\alpha}^T K_j \vec{\alpha}\rvert\rvert_\infty$
$\mathbf{L_1}$	$\theta_j = \bar{\theta}\ (\delta = 0)$	$\max \lvert\lvert\vec{\alpha}^T K_j \vec{\alpha}\rvert\rvert_1$
$\mathbf{L_2}$	$\lvert\lvert\theta_j\rvert\rvert_2 = 1\ (\delta = 2)$	$\max \lvert\lvert\vec{\alpha}^T K_j \vec{\alpha}\rvert\rvert_2$

- This extension can be applied to a wide range of kernel fusion algorithms

One class Support Vector Machine



One-Class SVM Classifier. The origin is the only original member of the second class.

(Tax and Duin, 1999)
(Schölkopf *et al.*, 2001)
(Manevitz and Yousef, 2001)

$$\boxed{\text{P:}}\ \min_{\vec{w},\xi,\rho}\ \frac{1}{2}\vec{w}^T\vec{w} - \frac{1}{\nu l}\sum_{k=1}^{l}\xi_k - \rho$$

$$\text{s.t.}\quad \vec{w}^T\phi(\vec{x}_k) \geq \rho - \xi_k,\ \ k = 1, \ldots, N$$

$$\xi_k \geq 0,\ \ k = 1, ..., N.$$

$\vec{w}$: the norm vector of the separating hyperplane
$\vec{x}_k$: the training samples
ν: a regularization term penalizing the outliers in the training samples
$\phi(\cdot)$: the feature map
ρ: the bias term
ξ_k: the slack variables
N: the number of training samples

$$\boxed{\text{D:}}\ \min_{\vec{\alpha}}\ \boxed{\vec{\alpha}^T K \vec{\alpha}}$$

$$\text{s.t.}\quad 0 \leq \alpha_k \leq \frac{1}{\nu N},\quad k = 1, ..., N$$

$$\sum_{k=1}^{N}\alpha_k = 1,$$

α_k: the dual variables
K: the kernel matrix

Kernel fusion in one class SVM

- L_∞-norm kernel fusion (De Bie *et al.*, 2007)

$$\begin{aligned} &\min_{\vec{\alpha}} \quad t \\ &\text{s.t.} \quad t \geq \vec{\alpha}^T K_j \vec{\alpha}, \;\; j = 1, ..., p \\ &\qquad 0 \leq \alpha_k \leq \frac{1}{\nu N}, \quad k = 1, ..., N \\ &\qquad \sum_{k=1}^{N} \alpha_k = 1, \end{aligned}$$

p: the number of kernel matrices

K_j: the j-th kernel matrix

- L_2-norm kernel fusion (Yu *et al.*, 2009)

$$\begin{aligned} &\min_{\vec{\alpha}} \quad t \\ &\text{s.t.} \quad t \geq ||s_j||_2, \;\; j = 1, ..., p \\ &\qquad s_j \geq \vec{\alpha}^T K_j \vec{\alpha}, \;\; j = 1, ..., p \\ &\qquad 0 \leq \alpha_k \leq \frac{1}{\nu N}, \;\; k = 1, ..., N \\ &\qquad \sum_{k=1}^{N} \alpha_k = 1. \end{aligned}$$

s_j: dummy variables

Disease gene prioritization

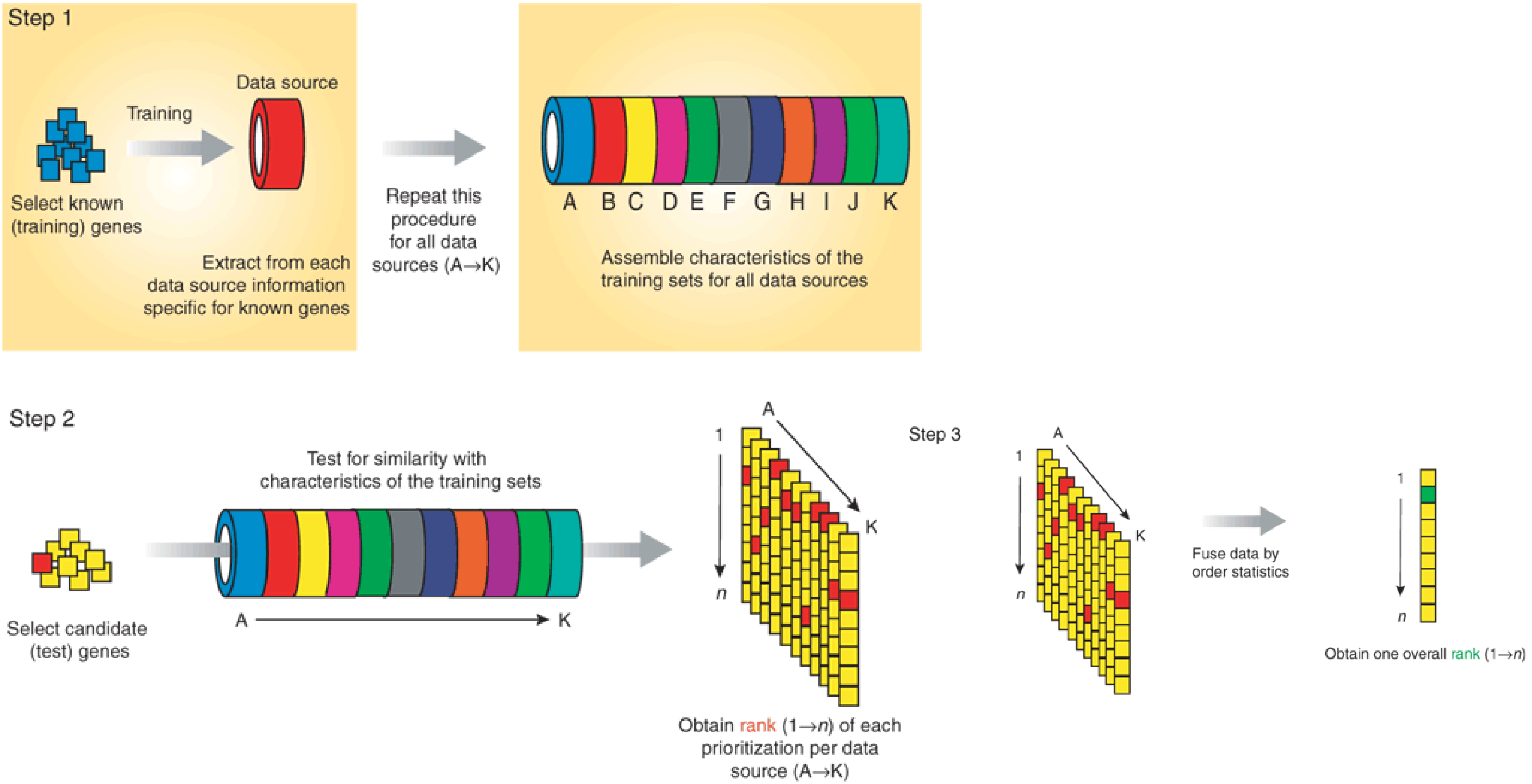
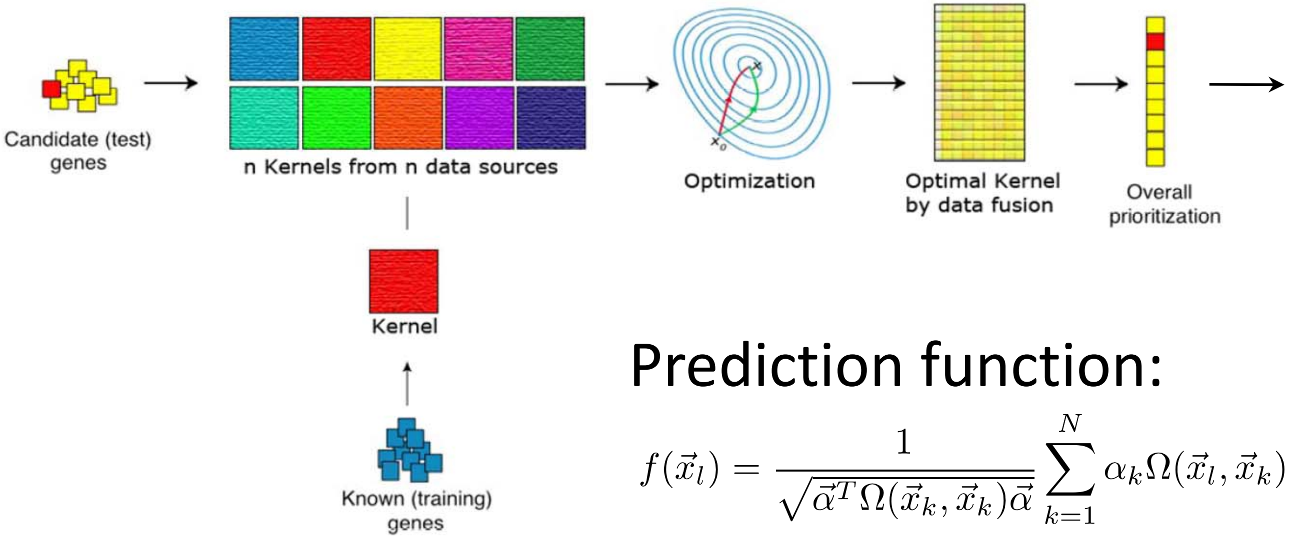
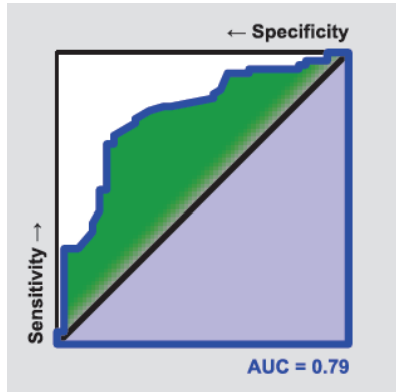


Image adapted from Aerts *et al.*, 2006

Case study 1: kernel based disease gene prioritization





Prediction function:

$$f(\vec{x}_l) = \frac{1}{\sqrt{\vec{\alpha}^T \Omega(\vec{x}_k, \vec{x}_k) \vec{\alpha}}} \sum_{k=1}^{N} \alpha_k \Omega(\vec{x}_l, \vec{x}_k)$$

$\{\vec{x}_k\}_{k=1}^{N}$: the training samples
$\vec{x}_l$: the candidate sample to be scored

ROC curve image adapted from http://www.svgopen.org/2008/papers/69-Evaluating_the_Quality_of_MultipleChoice_Tests_with_Automatically_Generated_Visualizations/sample_roc.png

Application 1: kernel based disease gene prioritization

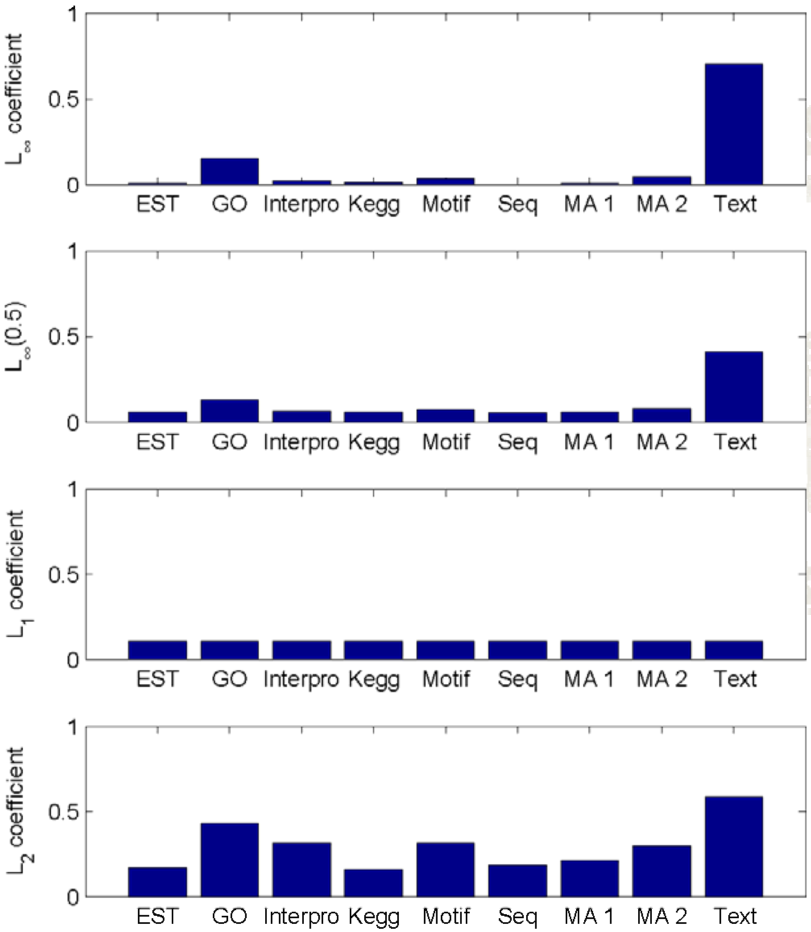
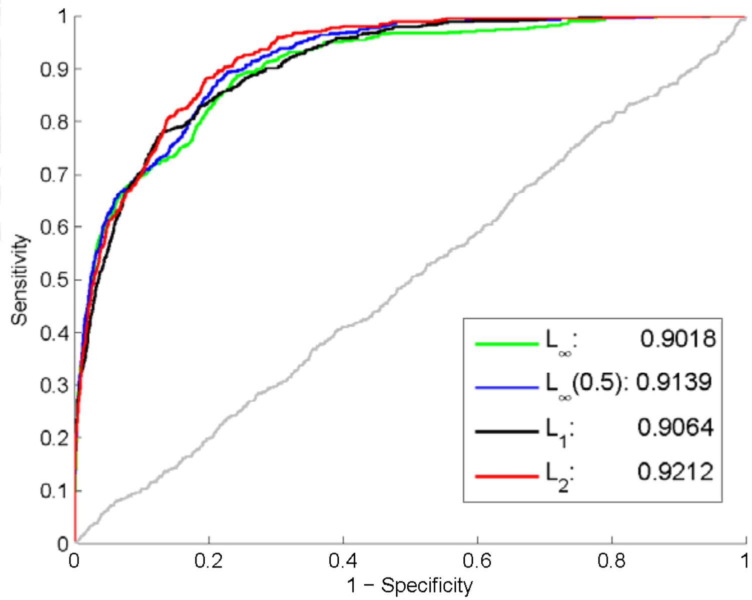


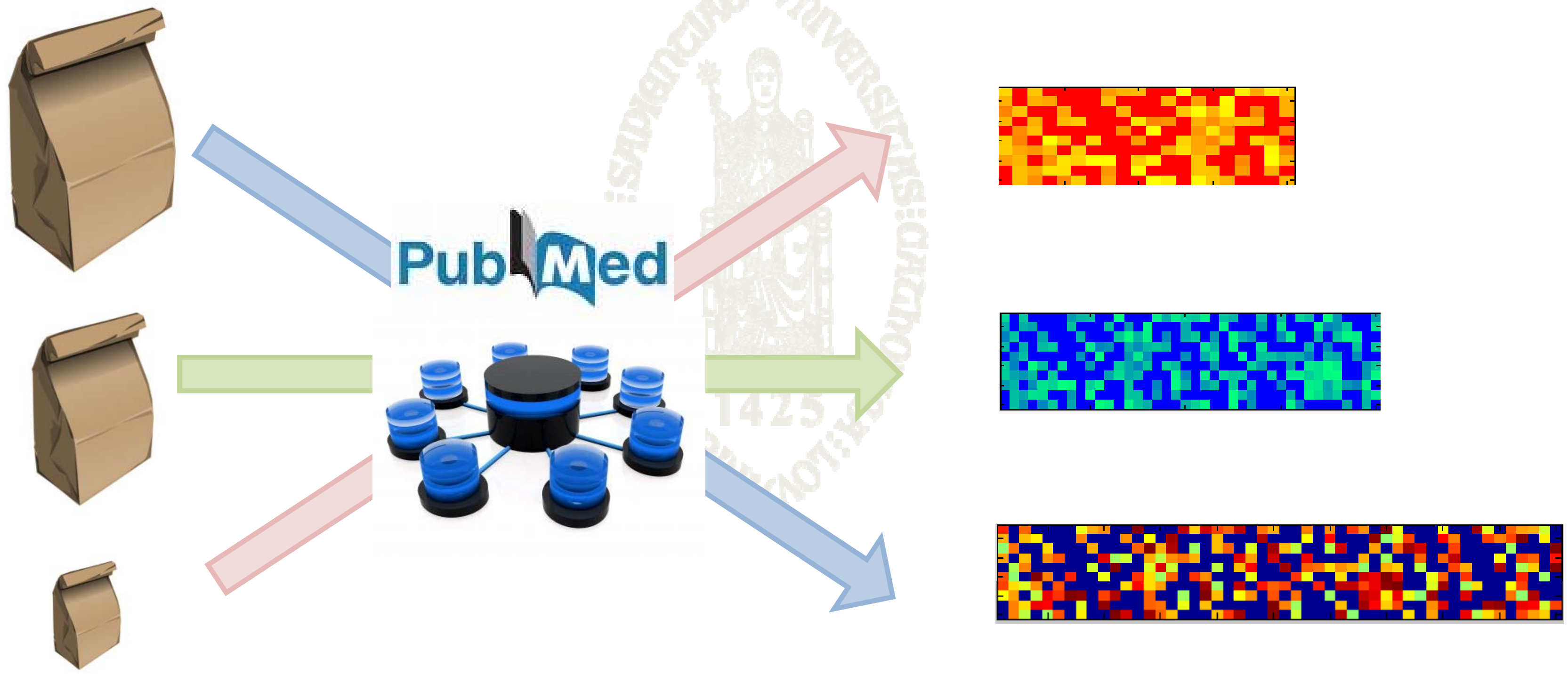
Table 1: AUC values of LOO performance evaluated from 20 random repetitions. The paired Spearman correlation scores indicate the similarities of rankings obtained by different approaches compared with the target rankings (denoted as -).

	AUC	corr	corr	corr	corr
L_∞	0.9045(0.0043)	-	0.94	0.66	0.82
$L_\infty(0.5)$	0.9176(0.0040)	0.94	-	0.82	0.92
L_1	0.9103(0.0035)	0.66	0.82	-	0.90
L_2	**0.9219(0.0034)**	0.82	0.92	0.90	-



Application 2: gene prioritization by multi-view text mining

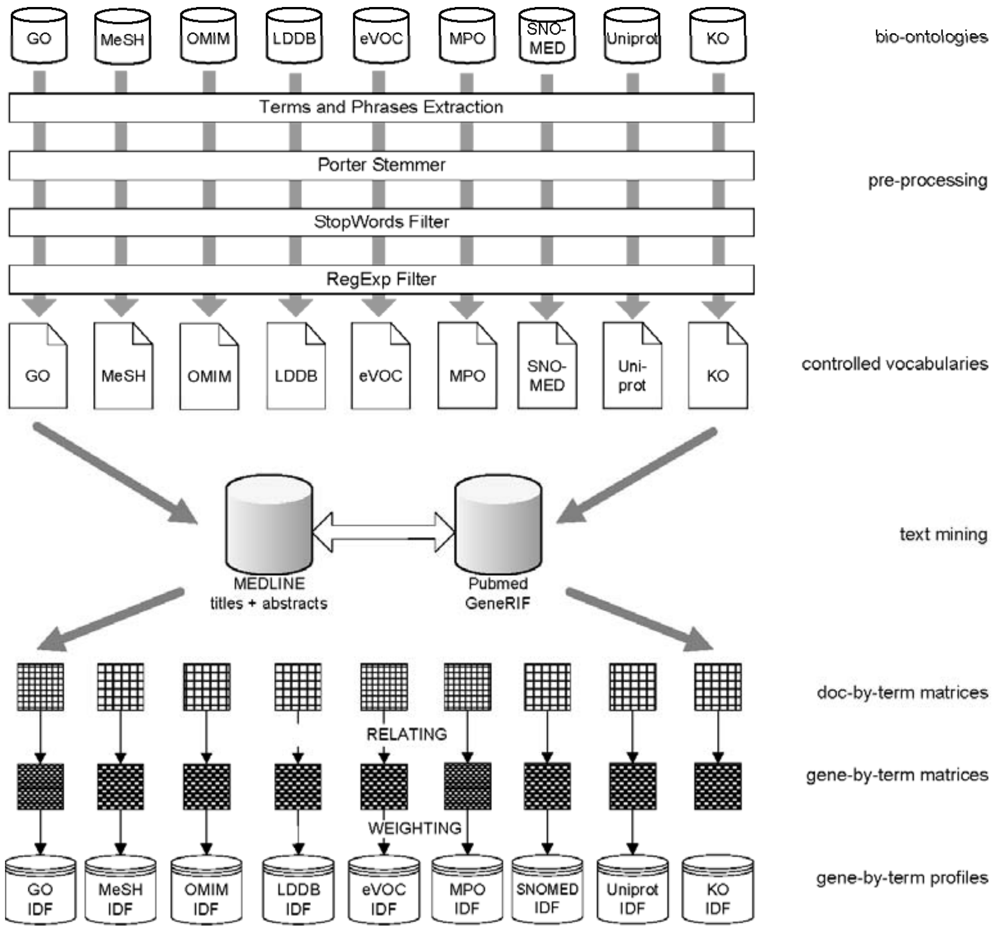
- Multi-view text mining



Bag of words

multi-view textual models

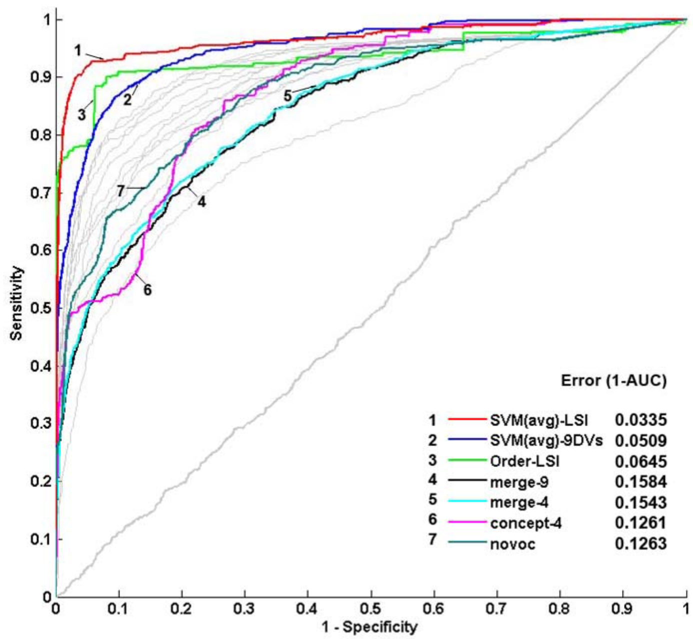
Application 2: gene prioritization by multi-view text mining



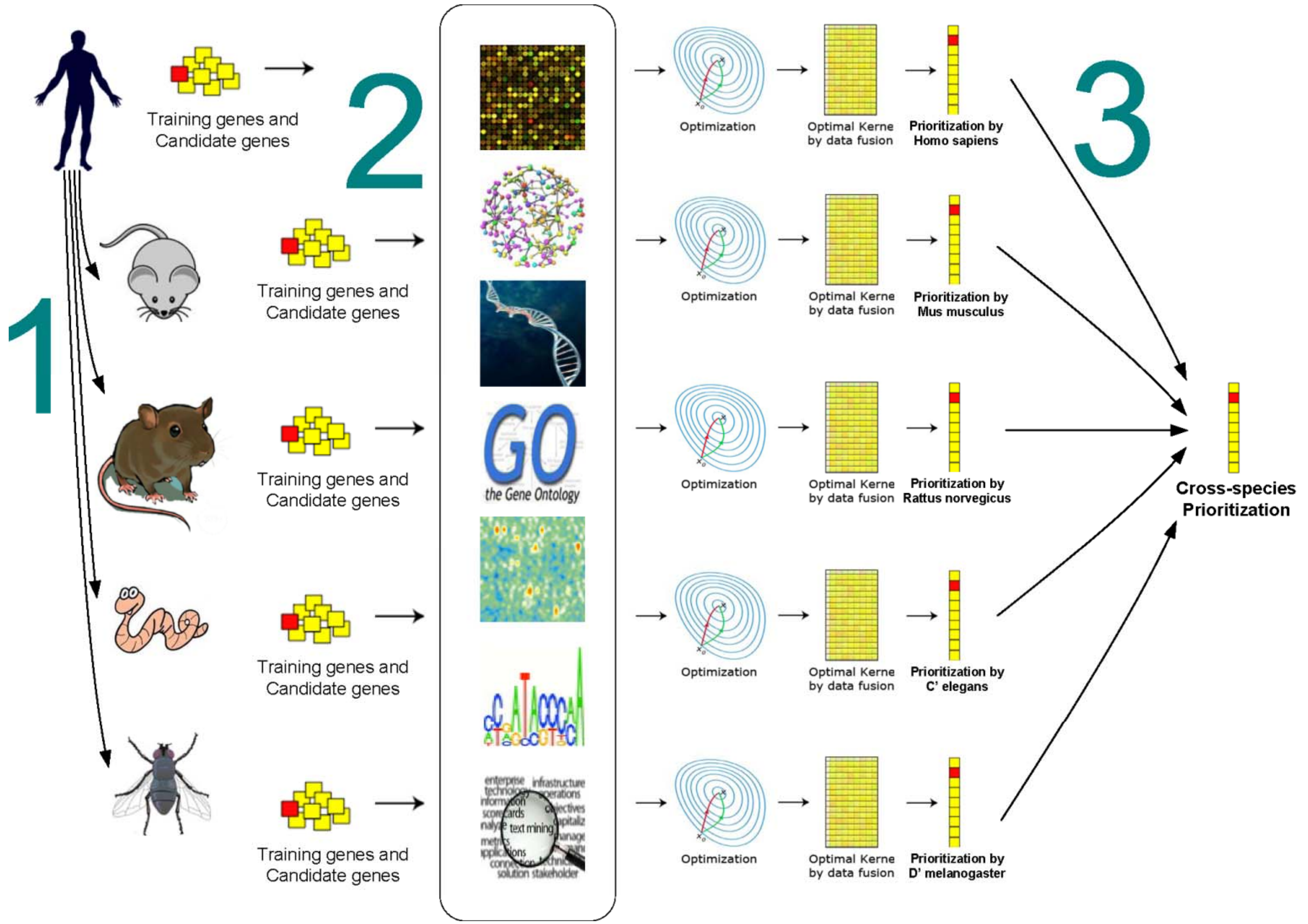
multi-view **text mining data**

Application 2: gene prioritization by multi-view text mining

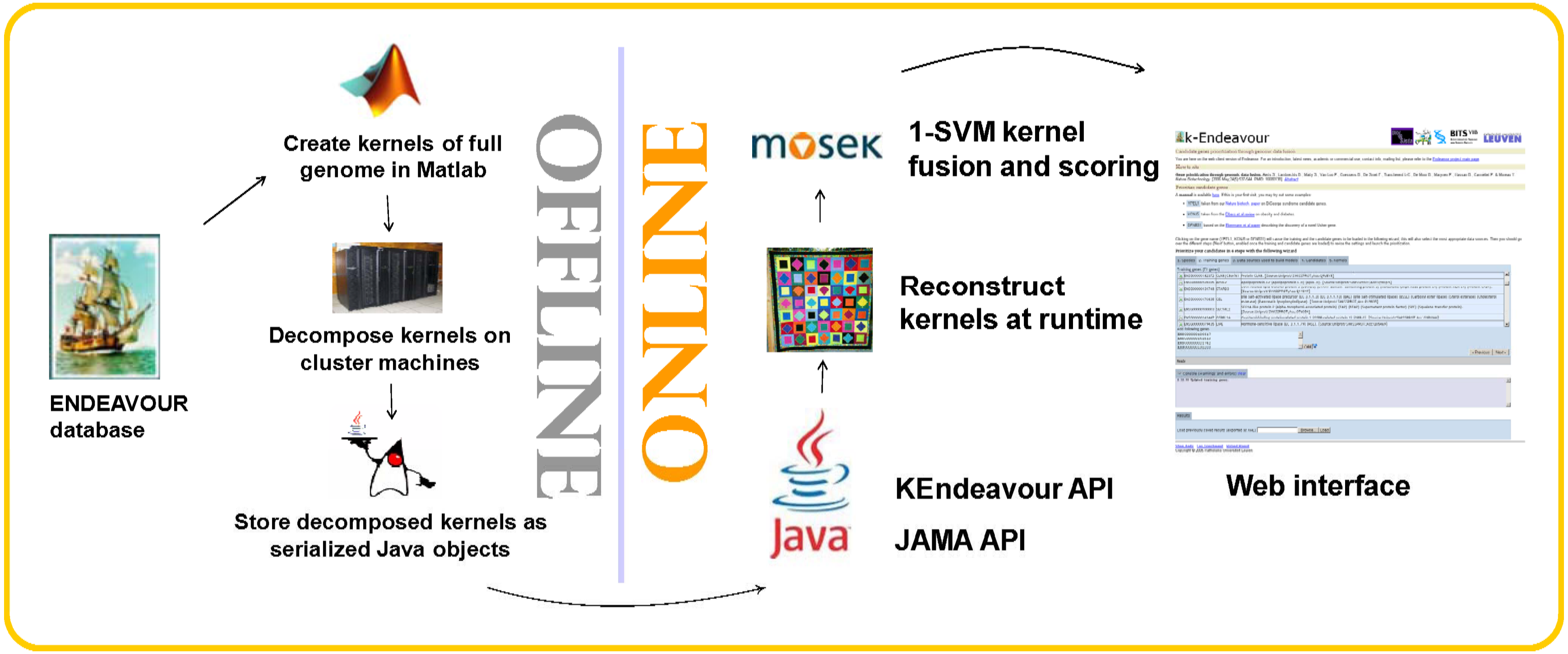
- Multi-view performs better than merging VOCs
- Multi-view performs better than the best individual VOC
- Kernel fusion + Latent semantic indexing performs the best



Software: Endeavour MerKator



Software: Endeavour MerKator



Summary: Kernel fusion for one class problem

- Dual problem of one class SVM = the fundamental form of kernel fusion
- Comparison of L_∞-norm with L_2-norm kernel fusion
- Multi-view text mining for disease gene prioritization
  - data fusion + dimensionality reduction
- Endeavour MerKator software
  - single organism → multiple organisms
  - text mining data → multiple genomic data sets

Related publications

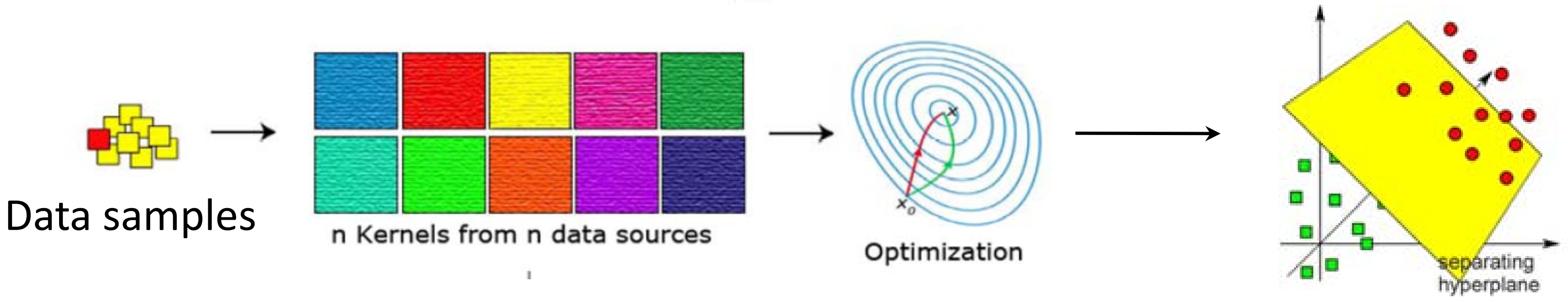
- **S. Yu**, S. Van Vooren, L.-C. Tranchevent, B. De Moor, Y. Moreau, "Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining", *Bioinformatics*, vol. 24, no. 16, pp. i119-125, 2008.

- **S. Yu**, L.-C. Tranchevent, B. De Moor, Y. Moreau, "Gene prioritization and clustering by multi-view text mining", *BMC Bioinformatics*, in publication, 2009.

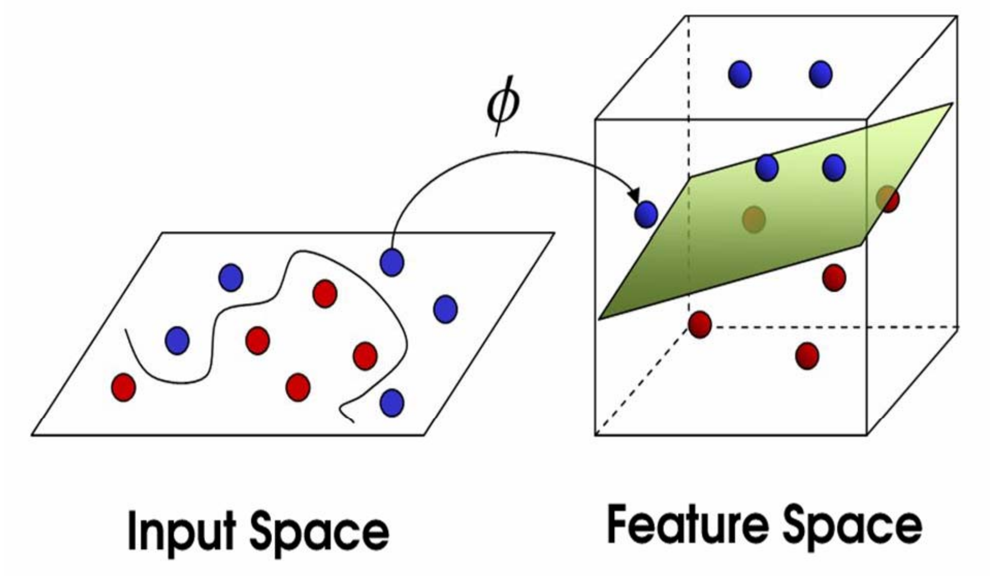
- L.-C.Tranchevent, R. Barriot, **S. Yu**, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, Y. Moreau, "ENDEAVOUR update: a web resource for gene prioritization in multiple species, *Nucleic Acids Research*, vol. 36, no. 1, pp. W377- W384, 2008.

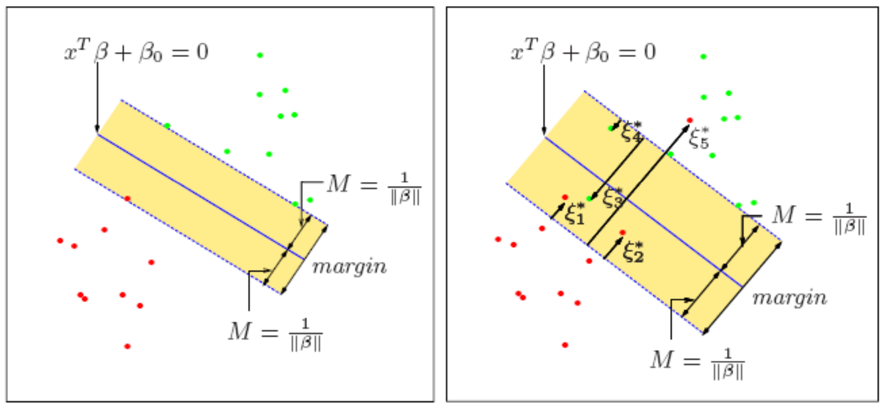
- **S. Yu**, L.-C. Tranchevent, S. Leach, R. Barriot, T. De Bie, B. De Moor, Y. Moreau, "Cross-species gene prioritization by genomic data fusion", *Manuscript in preparation*, 2009.

Topic 2: Kernel fusion for multi-class machine learning problems



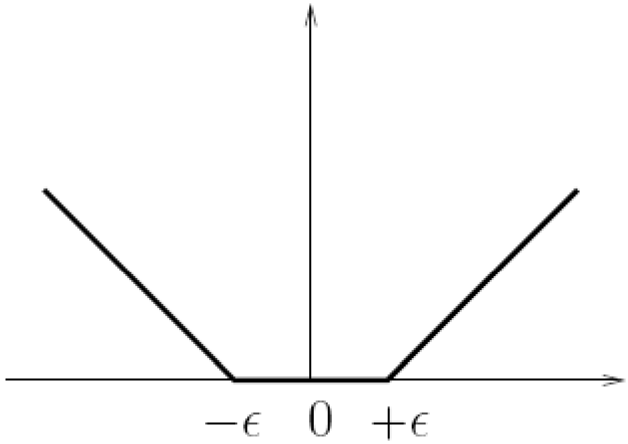
Support Vector Machines





Kernel fusion for multi-class problem

- Vapnik's SVM: hinge loss function



$$\boxed{\text{P:}} \quad \min_{\vec{w},b,\xi} \quad \frac{1}{2}\vec{w}^T\vec{w} + \lambda\sum_{k=1}^{N}\xi_k$$

$$\text{s.t. } y_k[\vec{w}^T\phi(\vec{x}_k) + b] \geq 1 - \xi_k, \; k = 1, ..., N$$

$$\xi_k \geq 0, \; k = 1, ..., N,$$

***L*$_\infty$-norm solution**:
(Lanckriet *et al.*, 2004; Bach *et al.*, 2003)

$$\boxed{\text{D:}} \quad \min_{\gamma,\vec{\alpha}} \quad \frac{1}{2}\gamma - \vec{\alpha}^T\vec{1}$$

$$\text{s.t. } (Y\vec{\alpha})^T\vec{1} = 0,$$

$$0 \leq \alpha_k \leq C, \; k = 1, ..., N$$

$$\gamma \geq \vec{\alpha}^T Y K_j Y \vec{\alpha}, \quad j = 1, ..., p,$$

***L*$_2$-norm solution**: (Yu *et al.*, 2009)

$$\boxed{\text{D:}} \quad \min_{\eta,\vec{\alpha}} \quad \frac{1}{2}\eta - \vec{\alpha}^T\vec{1}$$

$$\text{s.t. } (Y\vec{\alpha})^T\vec{1} = 0,$$

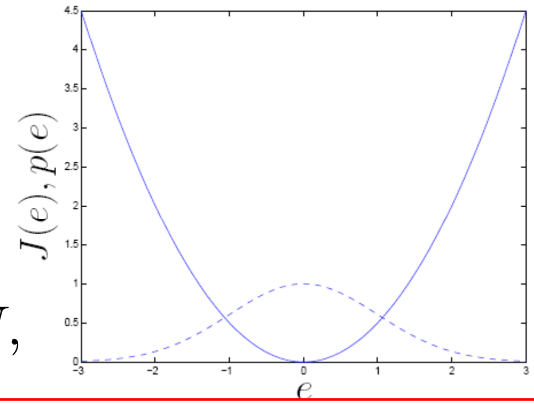
$$0 \leq \alpha_k \leq C, \quad k = 1, ..., N$$

$$\eta \geq ||\gamma_j||_2, \quad j = 1, ..., p$$

$$\gamma_j \geq \vec{\alpha}^T Y K_j Y \vec{\alpha}, \quad j = 1, ..., p.$$

Kernel fusion for multi-class problem

- Least squares SVM (LS-SVM)



$$\boxed{\text{P:}}\ \min_{\vec{w},b,\vec{e}} \ \frac{1}{2}\vec{w}^T\vec{w} + \frac{1}{2}\lambda\vec{e}^T\vec{e}$$

$$\text{s.t. } y_k[\vec{w}^T\phi(\vec{x}_k)+b] = 1 - e_k,\ \ k = 1, ..., N,$$

L_∞-norm solution:
(Ye *et al.*, 2008; Yu *et al.*, 2009)

$$\boxed{\text{D:}}\ \min_{\vec{\beta},t} \ \frac{1}{2}t + \frac{1}{2\lambda}\vec{\beta}^T\vec{\beta} - \vec{\beta}^T Y^{-1}\vec{1}$$

$$\text{s.t. } \sum_{k=1}^{N}\beta_k = 0,$$

$$t \geq \vec{\beta}^T K_j \vec{\beta},\ \ j = 1, ..., p.$$

L_2-norm solution: (Yu *et al.*, 2009)

$$\boxed{\text{D:}}\ \min_{\vec{\beta},\eta} \ \frac{1}{2}\eta + \frac{1}{2\lambda}\vec{\beta}^T\vec{\beta} - \vec{\beta}^T Y^{-1}\vec{1}$$

$$\text{s.t. } \sum_{k=1}^{N}\beta_k = 0,$$

$$s_j \geq \vec{\beta}^T K_j \vec{\beta},\ \ j = 1, ..., p,$$

$$\eta \geq ||s_j||_2,\ \ j = 1, ..., p.$$

Application 3: clinical decision support by medical data fusion

- clinical decision support by integrating microarray and proteomics data (Daemen *et al.*, 2009)
- rectal cancer diagnosis of 36 patients
- tissue and plasma samples were gathered at three time points
  - before treatment (T0)
  - at the early therapy treatment (T1)
  - and at the moment of surgery (T2)
- 4 linear kernel matrices to combine
  - tissue samples: 2 microarray data sets (MA T0 and MA T1)
  - plasma samples: 2 proteomics data sets (PT T0 and PT T1)

Application 3: clinical decision support by medical data fusion

Table 3: Overall results of patient classification. In LSSVM L_∞ and L_2, the λ was estimated jointly as kernel coefficient. In LSSVM L_1, λ was set to 1. In all SVM approaches, the C parameter of box constraint equated to 1. In the table, the row and column labels respectively represent the numbers of genes (g) and proteins (p) used to construct the kernels. To evaluate the AUC of LOO validation, we ignored the bias term b (as the implicit bias approach) because its value varied by each left out sample. In our problem, considering the bias term decreased the AUC performance. The performance was compared among six algorithms at the same number of genes and proteins, where the best values are represented in bold, the second best ones in italic. The best performance of all the feature selection results is underlined. The complete experimental results containing 26 different numbers of genes and 26 numbers of proteins is available at http://homes.esat.kuleuven.be/~syu/l2lssvm.html

	LSSVM(L_∞)					SVM(L_∞)				
	14p	15p	16p	17p	18p	14p	15p	16p	17p	18p
24g	0.9416	0.9481	0.9253	0.9188	0.9188	0.8669	0.8896	0.8669	0.8669	0.8636
25g	0.9610	0.9610	0.9481	0.9383	0.9351	0.8864	0.8896	0.8766	0.8799	0.8766
26g	0.9513	0.9513	0.9188	0.9156	0.9123	0.8734	0.8864	0.8766	0.8701	0.8636
27g	0.9383	0.9351	0.9188	0.9123	0.9058	0.8571	0.8636	0.8636	0.8669	0.8539
28g	0.9448	0.9513	0.9383	0.9253	0.9286	0.8571	0.8669	0.8669	0.8636	0.8604
	LSSVM(L_1)					SVM(L_1)				
	14p	15p	16p	17p	18p	14p	15p	16p	17p	18p
24g	**0.9513**	**0.9513**	**0.9318**	**0.9318**	0.9253	0.9253	0.9416	0.9286	**0.9318**	0.9253
25g	**0.9643**	**0.9675**	**0.9578**	**0.9545**	**0.9545**	0.9416	0.9481	0.9351	0.9286	0.9286
26g	**0.9643**	**0.9643**	**0.9545**	**0.9545**	**0.9545**	0.9416	0.9481	0.9318	0.9318	0.9318
27g	**0.9643**	**0.9643**	**0.9545**	**0.9513**	**0.9481**	0.9383	0.9416	0.9286	0.9318	0.9318
28g	**0.9578**	**0.9675**	**0.9513**	**0.9513**	**0.9481**	0.9416	0.9416	0.9351	0.9351	0.9318
	LSSVM(L_2)					SVM(L_2)				
	14p	15p	16p	17p	18p	14p	15p	16p	17p	18p
24g	*0.9448*	**0.9513**	*0.9253*	*0.9221*	**0.9286**	0.9091	0.9123	0.9026	0.9058	0.8994
25g	*0.9610*	*0.9610*	*0.9513*	*0.9448*	*0.9448*	0.9253	0.9351	0.9188	0.9156	0.9156
26g	*0.9610*	*0.9545*	*0.9448*	*0.9351*	*0.9351*	0.9253	0.9416	0.9188	0.9221	0.9221
27g	*0.9578*	*0.9513*	*0.9448*	*0.9416*	*0.9351*	0.9221	0.9188	0.9156	0.9188	0.9188
28g	*0.9545*	**0.9675**	**0.9513**	*0.9416*	*0.9448*	0.9188	0.9286	0.9188	0.9221	0.9188

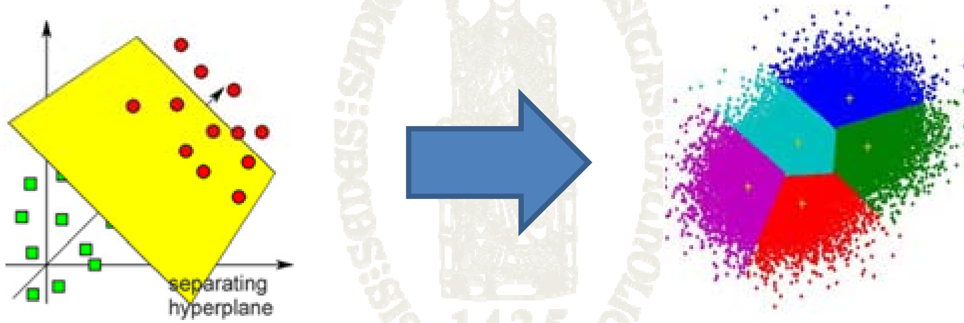
Summary: Kernel fusion for multi-class problem

- Main contributions: a novel L_2-norm LS-SVM formulation for multiple kernel learning
- LS-SVM becomes L_2 in dual aspects
  - primal: L_2-norm cost function (Suykens *et al.,* 2002)
  - dual: L_2-norm kernel fusion (Yu *et al.*, 2009)
- LS-SVM is also non-sparse in dual sets of variables
  - support vectors: non-sparse (Suykens *et al.,* 2002)
  - kernel coefficients: non-sparse (Yu *et al.*, 2009)
- In machine learning, the performance of an algorithm usually depends on the specific problem
- The computational efficiency is a solid advantage

Summary: Kernel fusion for multi-class problem

…

- From supervised learning to unsupervised learning



- Challenges
  - non-convex, NP-hard problem on unlabeled data
  - large scale data (training + test) and computational complexity
  - model evaluation and data collection

Topic 3: Kernel fusion for large scale data

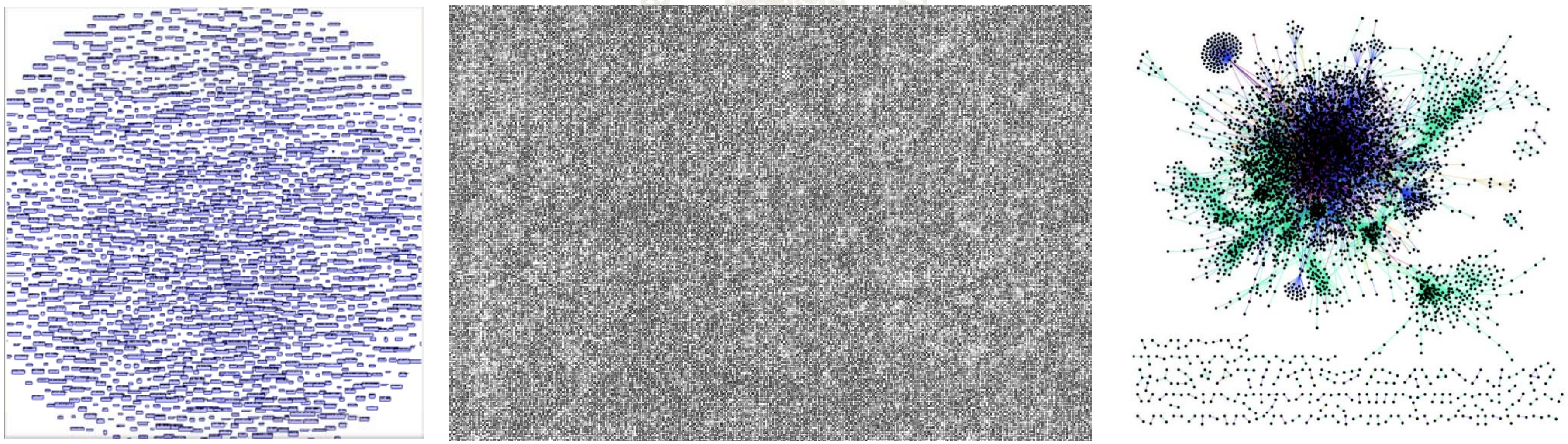
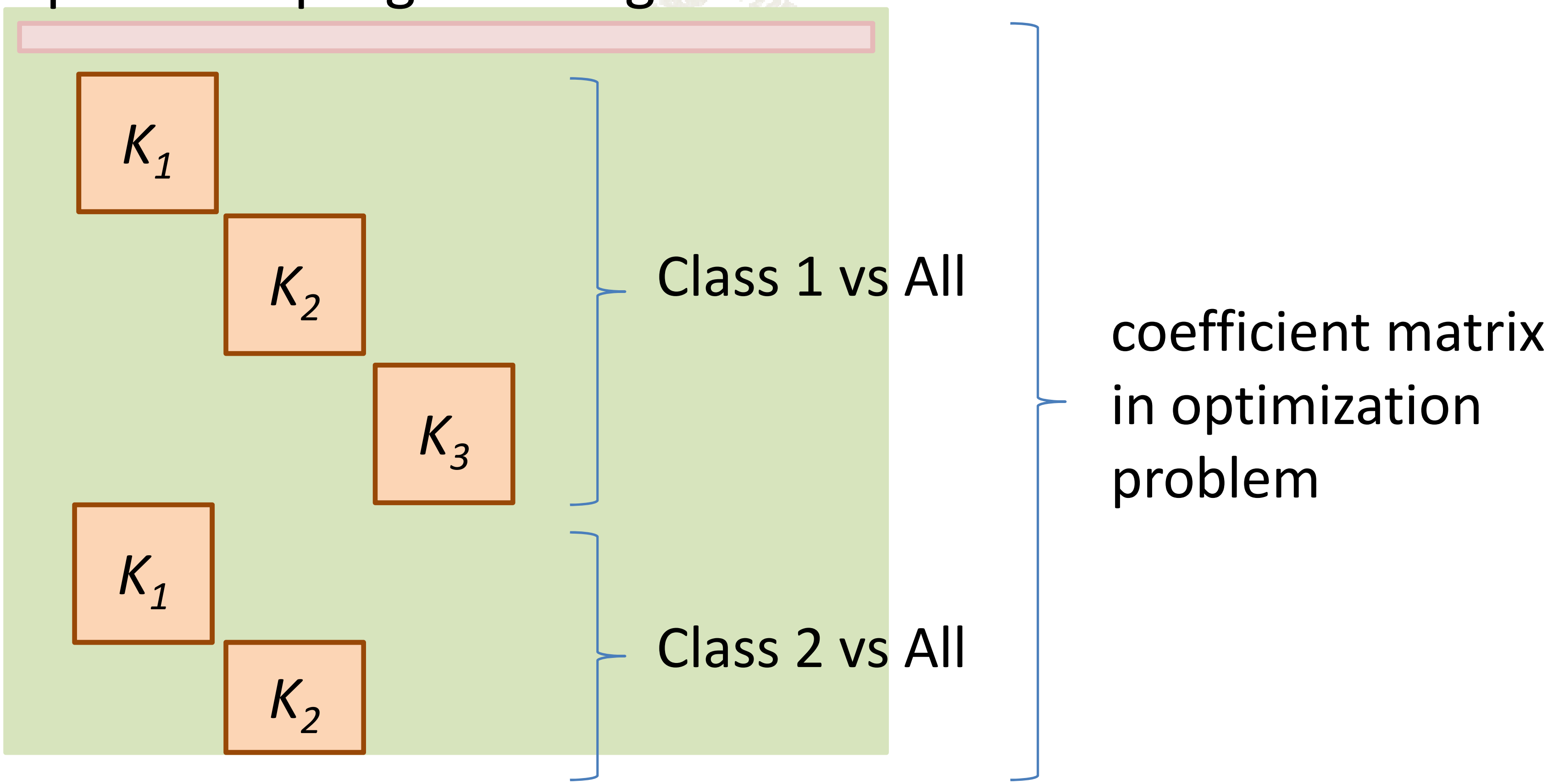


Figure adapted from http://sites.google.com/site/romainrigaux/hadoop-movies.png

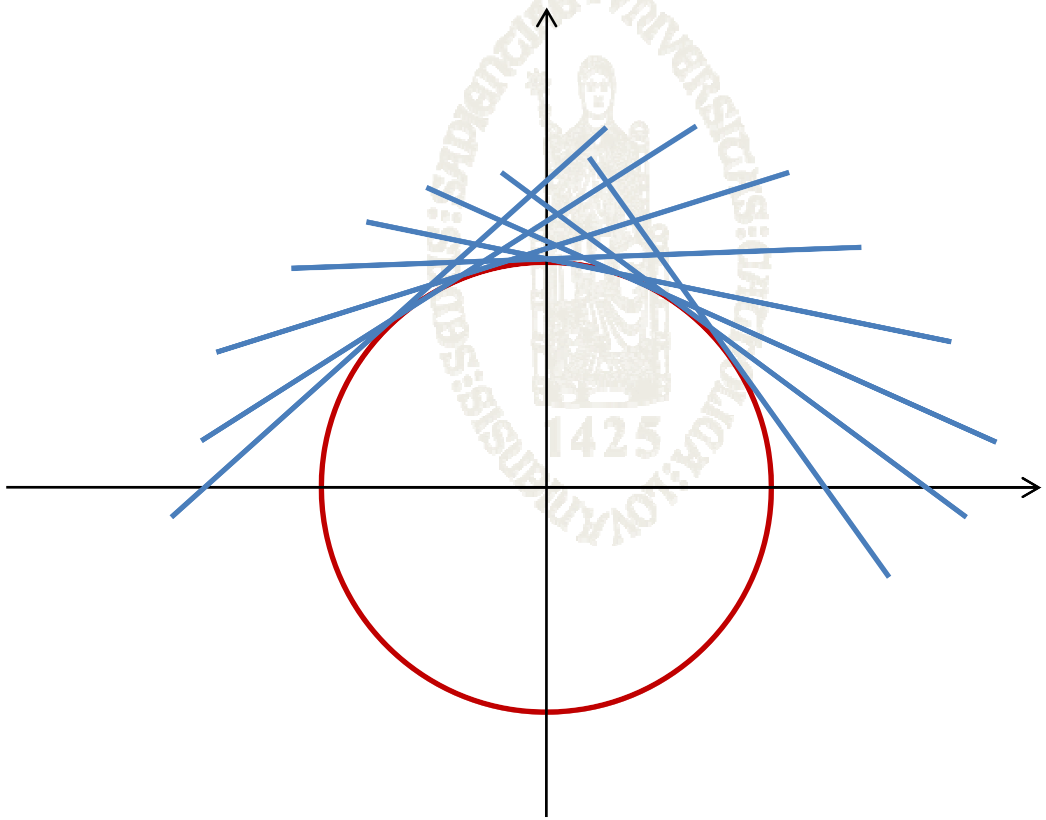
Computational burden of kernel fusion

- Memory intensive problem to solve kernel fusion as quadratic programming



Semi-infinite programming (SIP)

- A conceptual example



SIP for kernel fusion

- A Bi-level optimization approach

While there are violating constraints {

step1: to optimize the kernel coefficients

step2: to solve a single kernel SVM

}

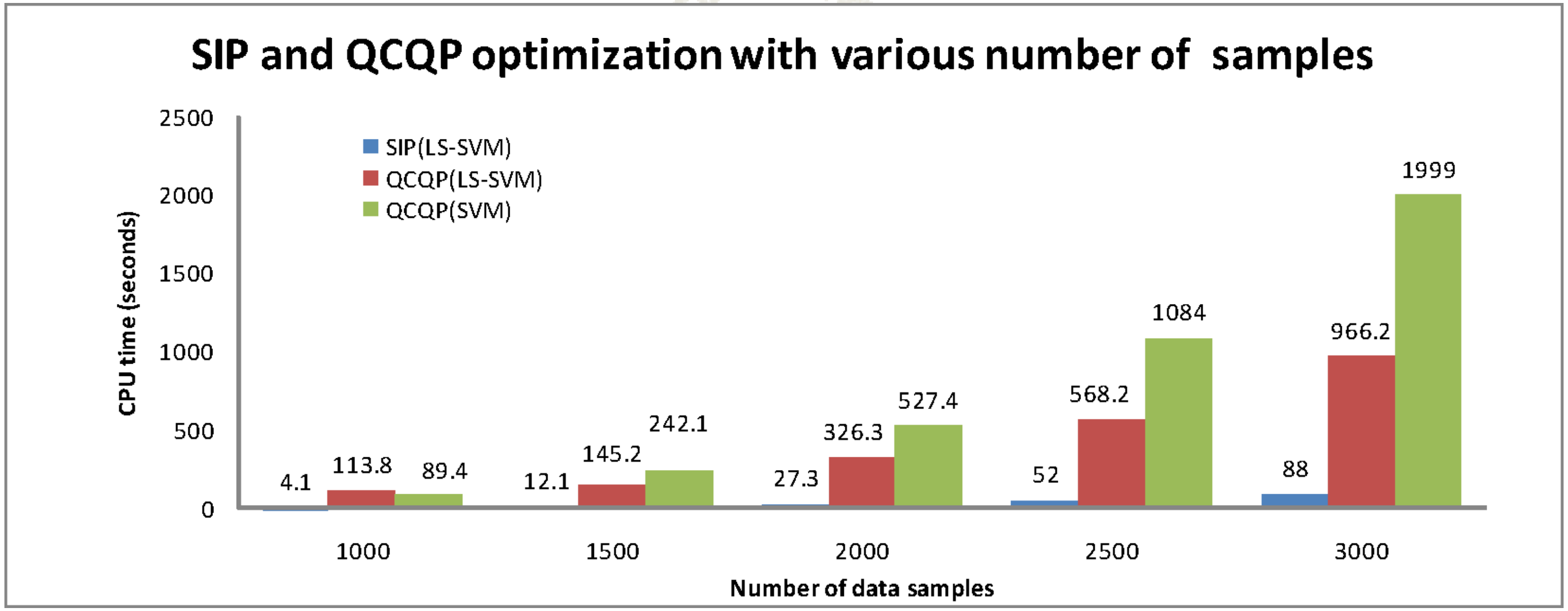
LS-SVM : a linear system solution!

$$\left[\begin{array}{c|c} 0 & \vec{1}^T \\ \hline \vec{1} & \Omega^{(\tau)} \end{array}\right] \left[\begin{array}{c} b^{(\tau)} \\ \hline \vec{\beta}^{(\tau)} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline Y^{-1}\vec{1} \end{array}\right],$$

where $\Omega^{(\tau)} = \sum_{j=1}^{p+1} \theta_j^{(\tau)} K_j$.

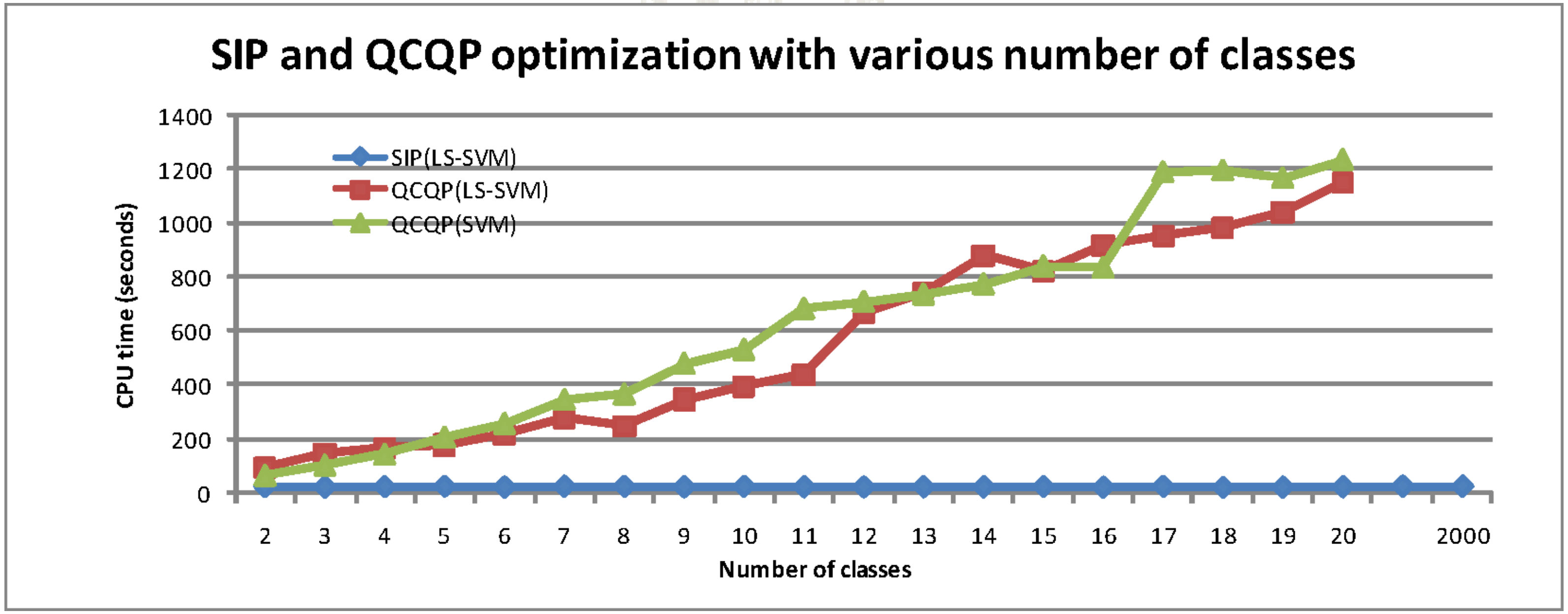
Efficiency of LSSVM kernel fusion

- Scale-up problem: the number of samples



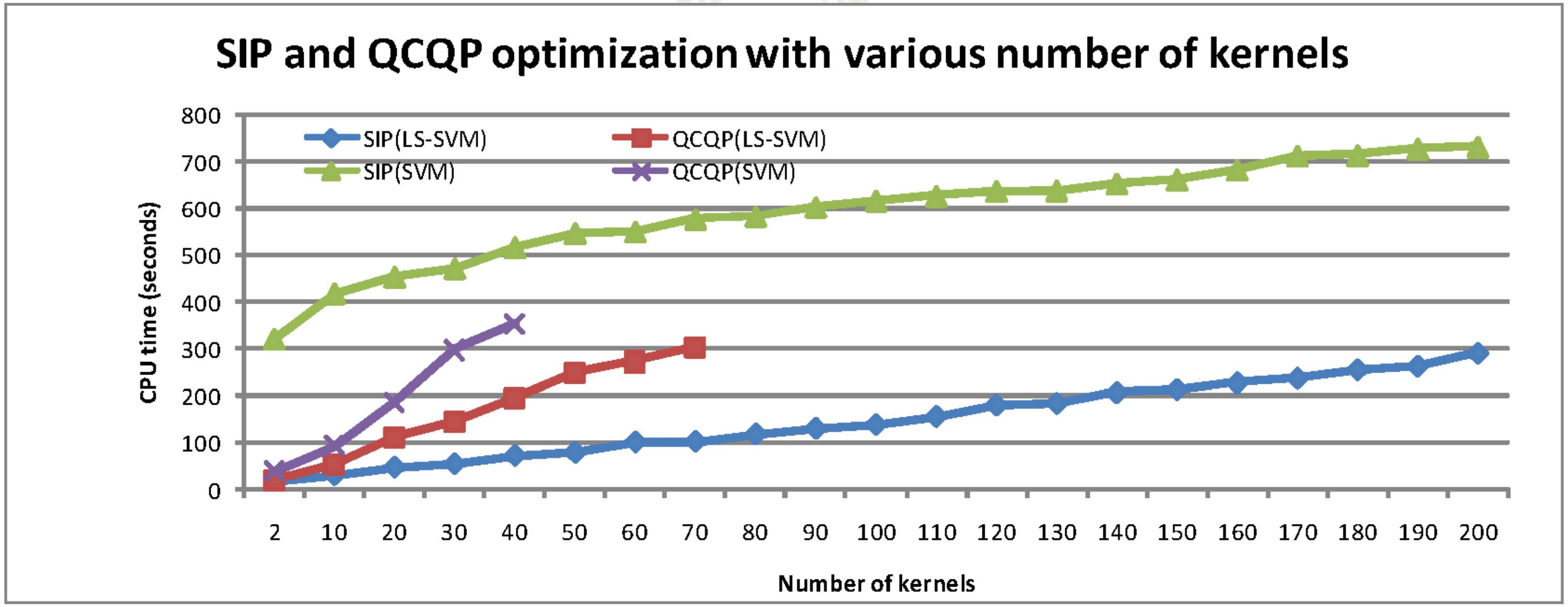
Efficiency of LSSVM kernel fusion

- Scale-up problem: the number of classes



Efficiency of LSSVM kernel fusion

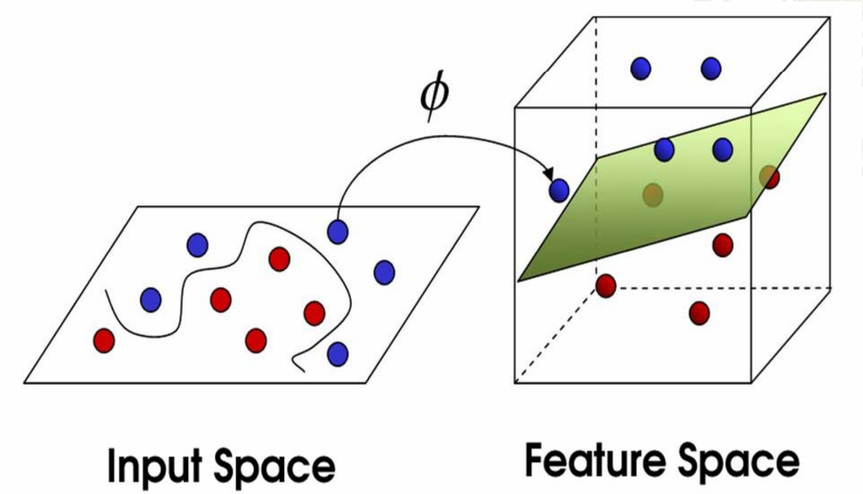
- Scale-up problem: the number of kernels

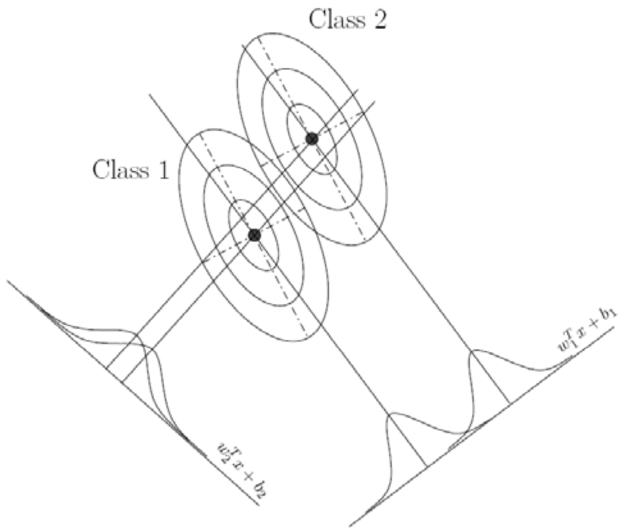


Rayleigh quotient objective of LSSVM kernel fusion

- The least squares problem is related to the Fisher Discriminant Analysis (Duda *et al.*, 2001)
- The LS-SVM is related to kernel Fisher Discriminant Analysis (Mika *et al.*, 1999; Suykens *et al.*, 2002)

$$\max_{\vec{w}} \mathcal{J}_{KFDA} = \frac{\vec{w}^T S_b^{\Phi} \vec{w}}{\vec{w}^T (S_t^{\Phi} + \kappa I) \vec{w}}$$





(Suykens *et al.*, 2002)

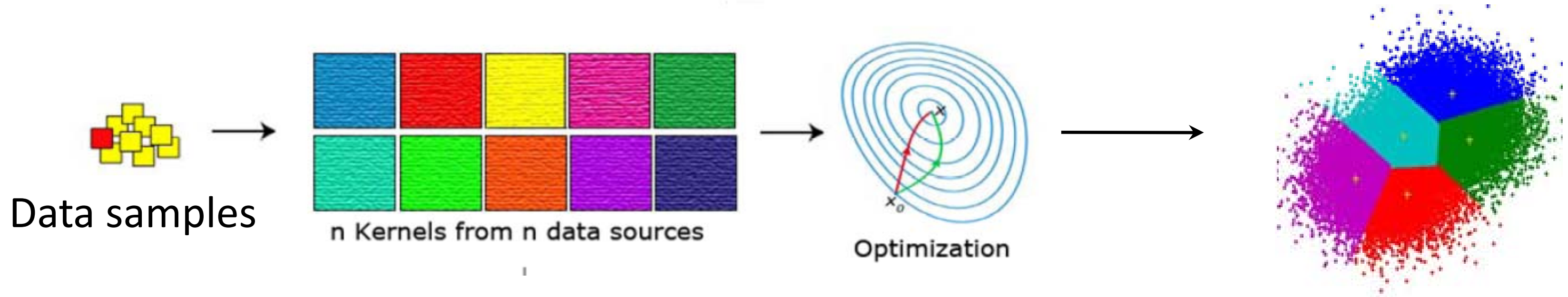
Summary of topic 3

- An efficient kernel fusion solution: SIP LS-SVM
- Solid advantages
- Rayleigh quotient objective of LS-SVM
  - Many unsupervised algorithms have the Rayleigh quotient objectives
  - LS-SVM can be plugged in unsupervised algorithms to create local optimal extensions for kernel fusion

Related publications (topic 2 & 3)

- **S. Yu**, T. Falck, A. Daemen, J. Suykens, B. De Moor and Y. Moreau, "Non-sparse kernel fusion and its applications in genomic data integration", *Internal Report, K.U.Leuven, Lirias number: 248322, submitted for publication*, 2009.

Topic 4: Kernel fusion for clustering analysis



Optimized data fusion for kernel K-means clustering (OKKC)

- An alternating minimization algorithm optimizing the Rayleigh quotient-like objective

$$
\boxed{\text{OKKC:}} \quad
\begin{aligned}
\max_{A,W,\vec{\theta}} \ \mathcal{J} = \ & \text{trace} \frac{S_b^{\Phi}}{\quad\quad}, \\
\text{s.t.} \quad & A^T A = I_k, \\
& W^T W = I_k, \\
& \mathbf{\Omega} = \sum_{r=1}^{p} \theta_r G_r, \\
& \theta_r \geq 0, \quad r = 1, ..., p \\
& \sum_{r=1}^{p} \theta_r = 1.
\end{aligned}
$$

KFDA (LSSVM)

A: the affinity cluster assignment matrix
W: the norm vector of separating hyperplane in KFDA
$\mathbf{\Omega}$: the combined kernel matrix
G_r: the r-th kernel matrix
θ_r: the kernel coefficients
p: the number of kernel matrices
S_b^{Φ}: between-cluster scatter matrix in kernel space

Optimized data fusion for kernel K-means clustering (OKKC)

Algorithm 0.1: $\text{OKKC}(G_1, G_2, ..., G_p, k)$

---

comment: Obtain the $\mathbf{\Omega}^{(\mathbf{0})}$ by the initial guess of $\theta_1^{(0)}, ..., \theta_p^{(0)}$

$A^{(0)} \leftarrow \text{Kernel K-means } (\mathbf{\Omega}^{(\mathbf{0})}, k)$

$\gamma = 0$

while $(\Delta A > \epsilon)$

$$\textbf{do} \begin{cases} step1 : F^{(\gamma)} \leftarrow A^{(\gamma)} \\ step2 : \mathbf{\Omega}^{(\gamma+1)} \\ \qquad \leftarrow \text{SIP-LS-SVM-MKL}(G_1, G_2, ..., G_p, F^{(\gamma)}) \\ step3 : A^{(\gamma+1)} \leftarrow \text{Kernel K-means}(\mathbf{\Omega}^{(\gamma+1)}, k) \\ step4 : \Delta A = ||A^{(\gamma+1)} - A^{(\gamma)}||^2 / ||A^{(\gamma+1)}||^2 \\ step5 : \gamma := \gamma + 1 \end{cases}$$

return $(A^{(\gamma)}, \theta_1^{(\gamma)}, ..., \theta_p^{(\gamma)})$

Performance of OKKC

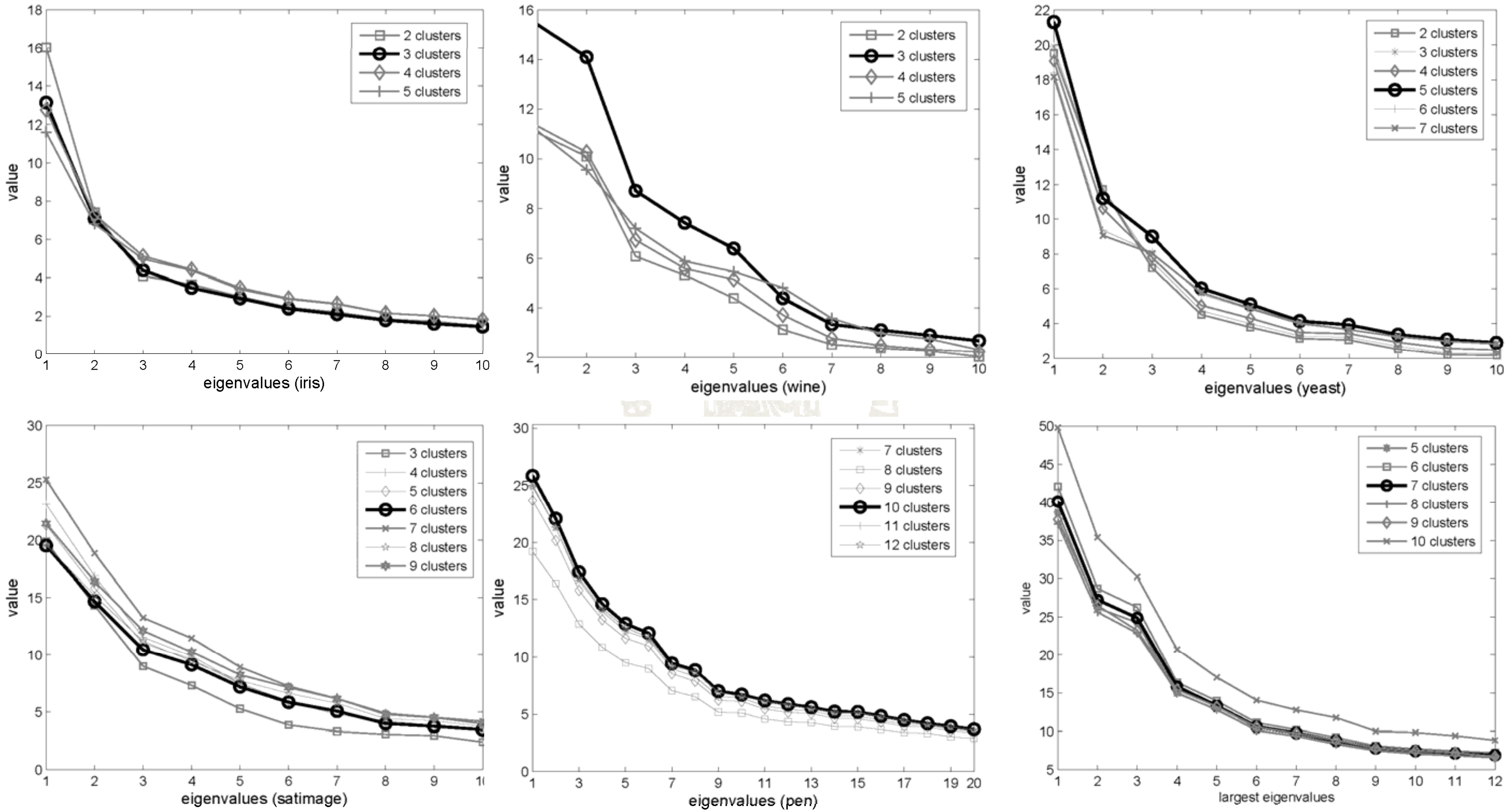
- Comparable performance to other algorithms
- More efficient (bi-level optimization, less memory requirement)

TABLE II

OVERALL RESULTS OF CLUSTERING PERFORMANCE

	best individual		worst individual		average combine			OKKC				NAML			
	ARI	NMI	ARI	NMI	ARI	NMI	time(sec)	ARI	NMI	itr	time(sec)	ARI	NMI	itr	time(sec)
Iris	0.7302 (0.0690)	0.7637 (0.0606)	0.6412 (0.1007)	0.7047 (0.0543)	0.7132 (0.1031)	0.7641 (0.0414)	0.22 (0.13)	0.7516 (0.0690)	0.7637 (0.0606)	7.8 (3.7)	5.32 (2.46)	0.7464 (0.0207)	0.7709 (0.0117)	9.2 (2.5)	15.45 (6.58)
Wine	0.3489 (0.0887)	0.3567 (0.0808)	0.0387 (0.0175)	0.0522 (0.0193)	0.3188 (0.1264)	0.3343 (0.1078)	0.25 (0.03)	0.3782 (0.0547)	0.3955 (0.0527)	10 (4.0)	18.41 (11.35)	0.2861 (0.1357)	0.3053 (0.1206)	6.7 (1.4)	16.92 (3.87)
Yeast	0.4246 (0.0554)	0.5022 (0.0222)	0.0007 (0.0025)	0.0127 (0.0038)	0.4193 (0.0529)	0.4994 (0.0271)	2.47 (0.05)	0.4049 (0.0375)	0.4867 (0.0193)	7 (1.7)	81.85 (14.58)	0.4256 (0.0503)	0.4998 (0.0167)	10 (2)	158.20 (30.38)
Satimage	0.4765 (0.0515)	0.5922 (0.0383)	0.0004 (0.0024)	0.0142 (0.0033)	0.4891 (0.0476)	0.6009 (0.0278)	4.54 (0.07)	0.4996 (0.0571)	0.6004 (0.0415)	10.2 (3.6)	213.40 (98.70)	0.4911 (0.0522)	0.6027 (0.0307)	8 (0.7)	302 (55.65)
Pen digit	0.5818 (0.0381)	0.7169 (0.0174)	0.2456 (0.0274)	0.5659 (0.0257)	0.5880 (0.0531)	0.7201 (0.0295)	15.95 (0.08)	0.5904 (0.0459)	0.7461 (0.0267)	8 (4.38)	396.48 (237.51)	0.5723 (0.0492)	0.7165 (0.0295)	8 (4.2)	1360.32 (583.74)
Disease genes	0.7585 (0.0043)	0.5281 (0.0078)	0.5900 (0.0014)	0.1928 (0.0042)	0.7306 (0.0061)	0.4702 (0.0101)	931.98 (1.51)	0.7641 (0.0078)	0.5395 (0.0147)	5 (1.5)	1278.58 (120.35)	0.7310 (0.0049)	0.4715 (0.0089)	8.5 (2.6)	3268.83 (541.92)
Journal sets	0.6644 (0.0878)	0.7203 (0.0523)	0.5341 0.0580	0.6472 0.0369	0.6774 (0.0316)	0.7458 (0.0268)	63.29 (1.21)	0.6812 (0.0602)	0.7420 (0.0439)	8.2 (4.4)	1829.39 (772.52)	0.6294 (0.0535)	0.7108 (0.0355)	9.1 (6.1)	4935.23 (3619.50)

Find clusters by OKKC



Optimized kernel Laplacian clustering (OKLC)

- To combine *heterogeneous* data structures
- Combination of attribute-based data (kernels) and interaction-based graphs (Laplacians)
- OKLC-$_{light}$: Bi-level optimization as OKKC using $\hat{L} = D^{-1/2} W D^{-1/2}$

$$\max_A \operatorname{trace} \left(A^T \hat{L} A + A^T X^{\Phi T} X^{\Phi} A \right)$$
$$\text{s.t.} \quad A^T A = I_k .$$

- OKLC: Tri-level optimization using $\tilde{L} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$

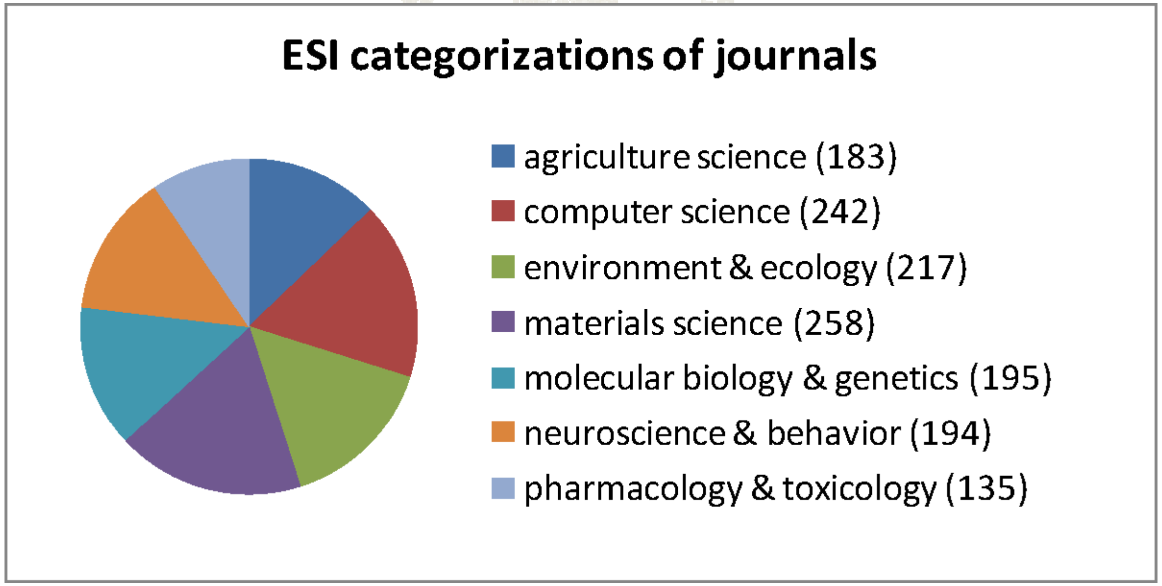
$$\max_A \operatorname{trace} \frac{A^T \mathbf{\Omega} A}{A^T \tilde{\mathbf{L}} A}$$
$$\text{s.t.} \quad A^T A = I_k .$$

Application 4: integrate lexical/citation information in journal set analysis

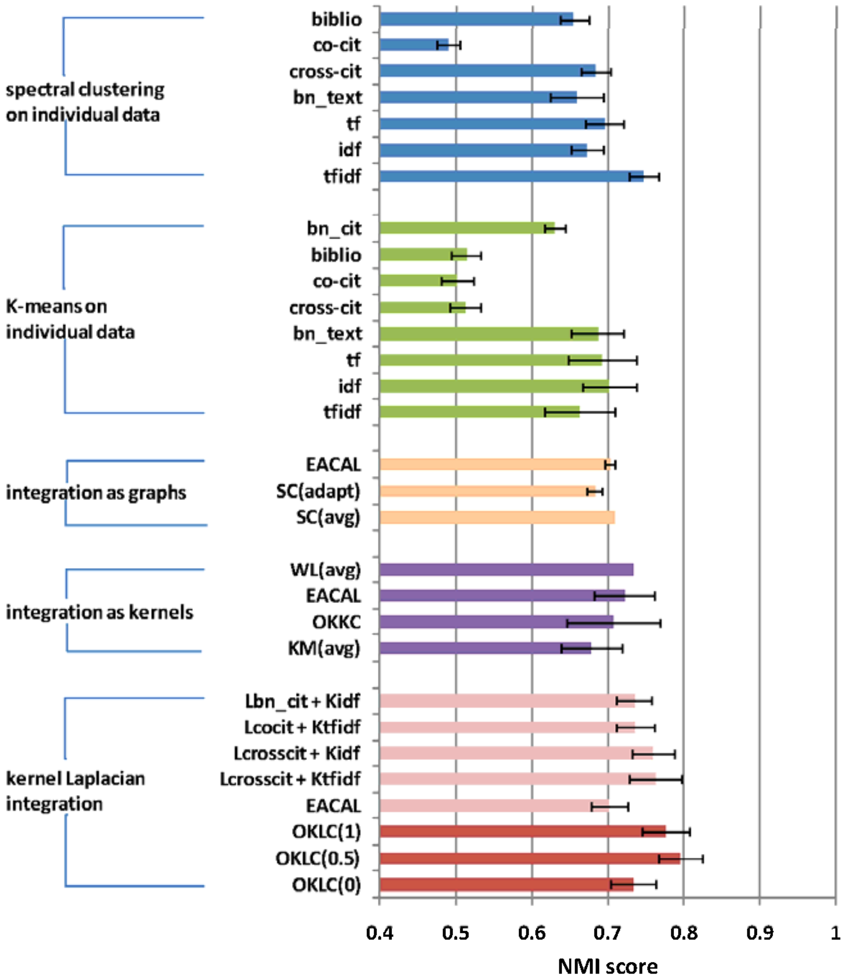
- Web of Science (WoS) database
- Papers of 1424 journals published from 2002 to 2006
- Text mining analysis (lexical similarity of journals)
  - no-controlled vocabulary → Zipf cut
  - 669,860 terms
  - TF-IDF, IDF, TF and binary weighting of terms
- Citation analysis (interaction of journals)
  - cross-citation
  - binary cross-citation
  - co-citation
  - bibliographic coupling

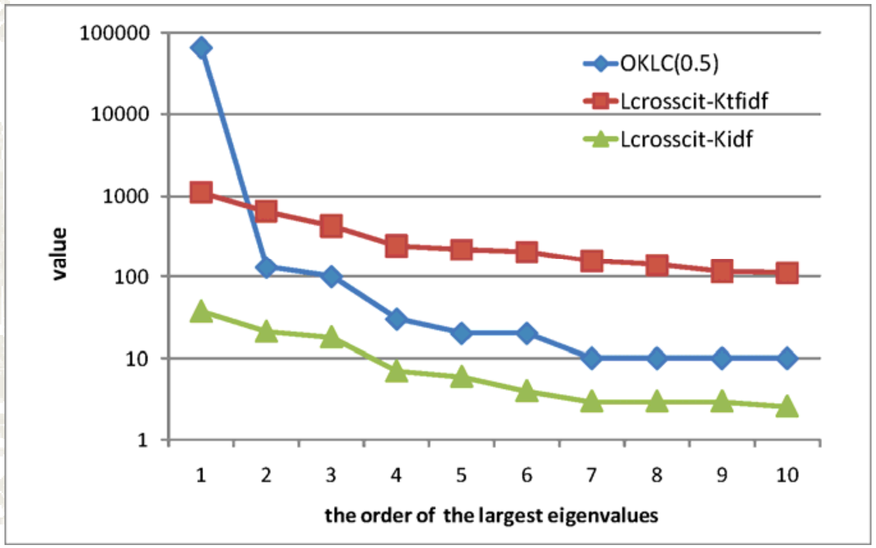
Application 4: integrate lexical/citation information in journal set analysis

- Combination of four linear kernels (lexical) and four Laplacians (citation)
- Evaluated by Essential Science Index (ESI)



Application 4: integrate lexical/citation information in journal set analysis

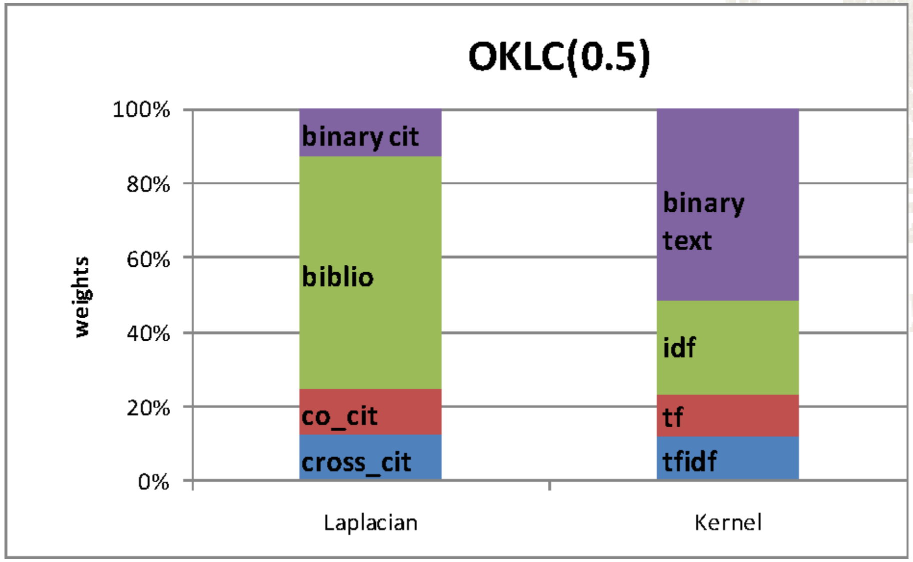


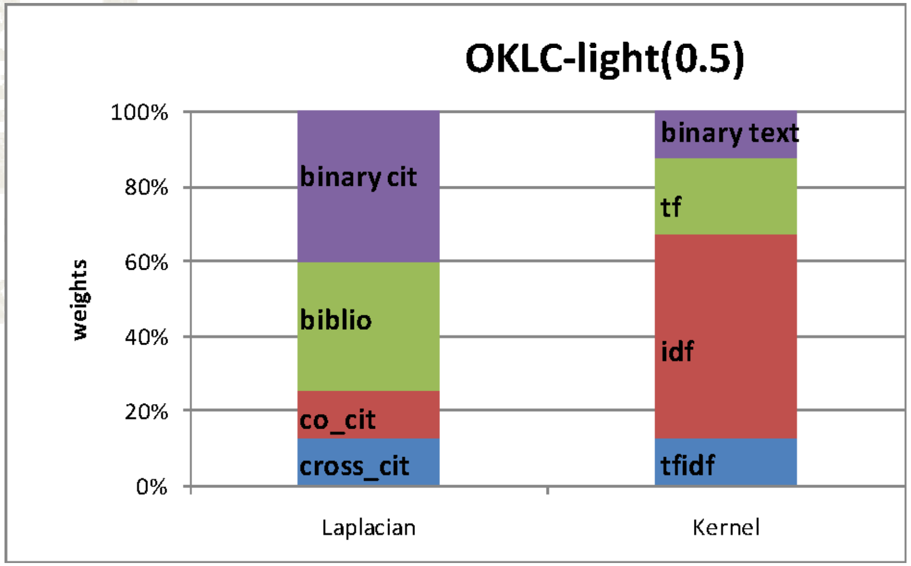


Comparison of OKLC and OKLC-$_{\text{light}}$

Table 11.1: Comparison of the OKLC and the OKLC-*light* on journal set clustering

	OKLC	OKLC-*light*
$\theta_{min} = 0$	0.7331 ± 0.0294	$\mathbf{0.7637} \pm 0.036$
$\theta_{min} = 0.5$	$\mathbf{0.7954} \pm 0.0285$	0.7734 ± 0.0367
$\theta_{min} = 1$	0.7762 ± 0.032	0.7726 ± 0.0156





Summary of topic 4

- Preliminary efforts to incorporate kernel fusion with unsupervised learning
- Combination of heterogeneous data sources (OKKC)
- Combination of heterogeneous data structures (OKLC)
- Many remaining challenges / opportunities
  - statistical validation of clustering in data fusion
  - multi-objective clustering
  - clustering with overlapping memberships
  - stability issue / sampling technique
  - kernel evaluation without validation

Related publications

- **S. Yu**, L.-C. Tranchevent, X. Liu, W. Glänzel, J. Suykens, B. De Moor and Y. Moreau, "Optimized data fusion for kernel K-means clustering", *Internal Report 08-200, ESAT-SISTA, K.U.Leuven, 2008, Lirias number: 242275, submitted for publication.*

- **S. Yu**, X. Liu, W. Glänzel, B. De Moor, Y. Moreau, "Clustering with multiple kernels and Laplacians", *Internal report, K.U.Leuven, submitted for publication*, 2009.

- X. Liu, **S. Yu**, Y. Moreau, B. De Moor, W. Glänzel, F. Janssens, "Hybrid clustering of text mining and bibliometrics applied to journal sets", *in Proceeding of SIAM Data Mining (SDM) Conference*, 2009. **(equally contributed author)**

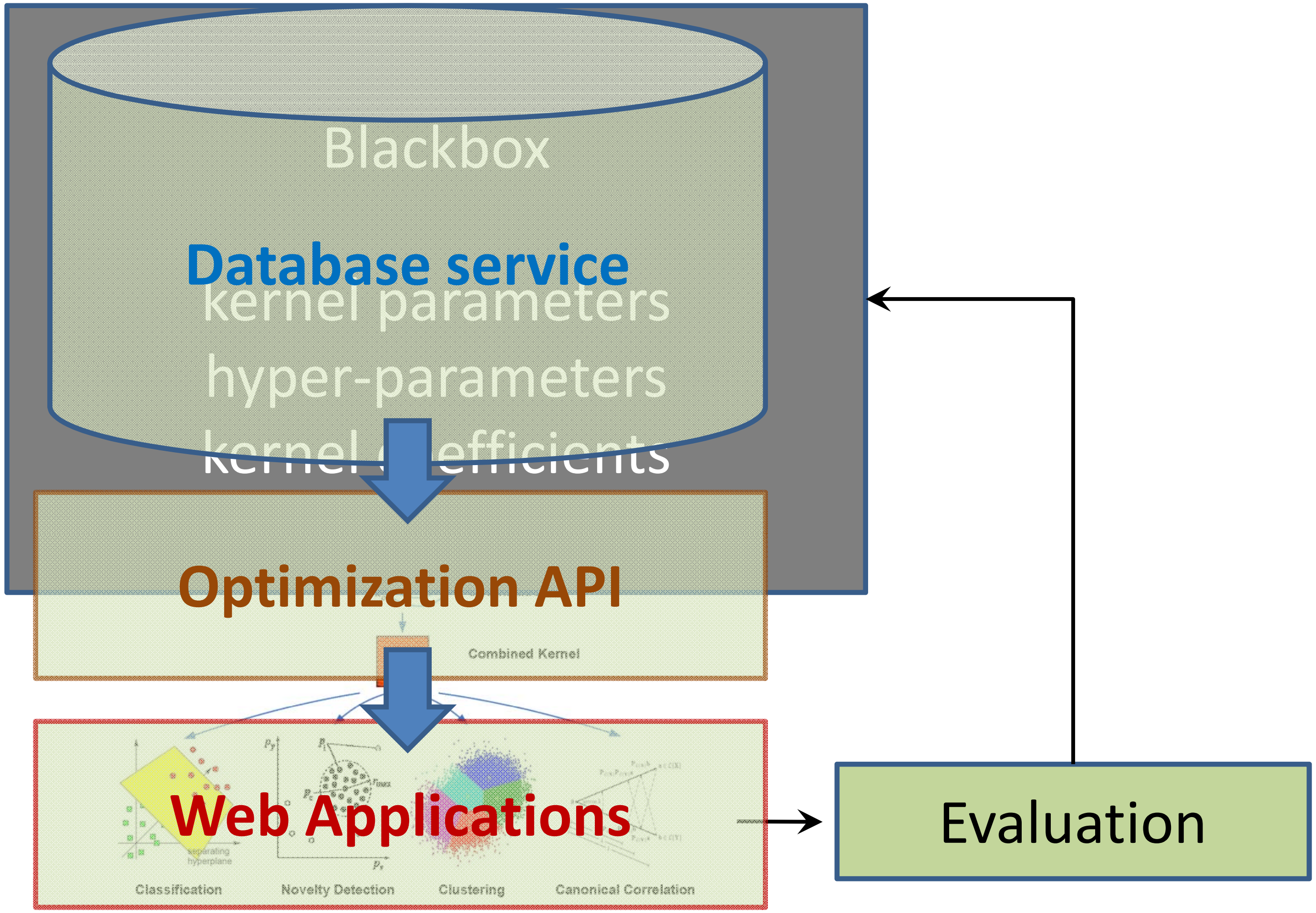
- **S. Yu**, B. De Moor, Y. Moreau, "Clustering by heterogeneous data fusion: framework and applications", *in Workshop of learning from multiple data sources, NIPS*, 2008.

- X. Liu, **S. Yu**, Y. Moreau Y, B. De Moor, W. Glänzel, F. Janssens, "Hybrid Clustering by Integrating Text and Citation based Graphs in Journal Database Analysis", *accepted for publication in ICDM-09 Workshop on Mining Multiple Information Sources*, 2009.

Conclusions and Future Research

Figure adapted from www.clipart.com

Kernel-based data fusion: a view not only for bioinformatics



Kernel-based data fusion: a unified model

$$\mho = f\left\{\mathbb{P}, \mathbb{N}, \mathbb{S}, \mathbb{A}, \mathbb{O}, \cdots\right\}$$

Theory

Algorithm

Application

Software

one-class
multi-class
clustering
canonical correlation
regression
…

L_∞-norm
L_1-norm
L_2-norm
L_p-norm
…

kernels
Laplacians
multi-objective
…

Bioinformatics
Text mining
Robotics
Image and Signal processing
Scientometrics
…

Convex optimization

Stochastic optimization

High-order SVD (Tensor)
…

Interesting topics for future researches

- Kernel-based sensor fusion
- Bioinformatics: integration of expression data and interaction network
- A joint framework of data fusion with feature selection
- Non-additive models for kernel fusion

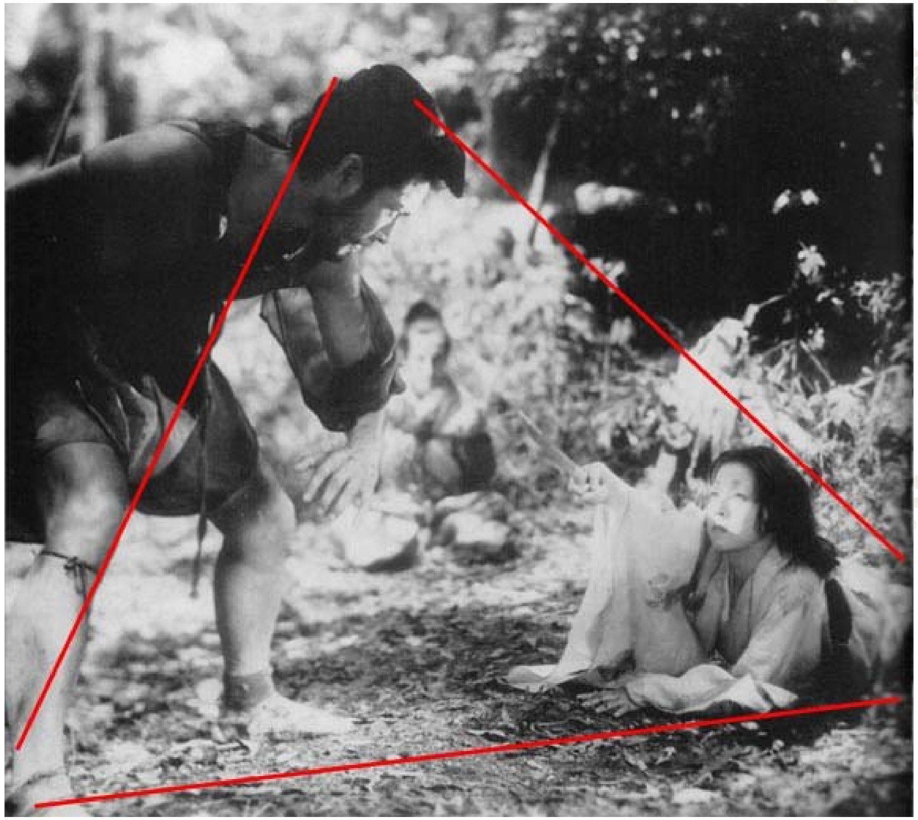
Closing remarks

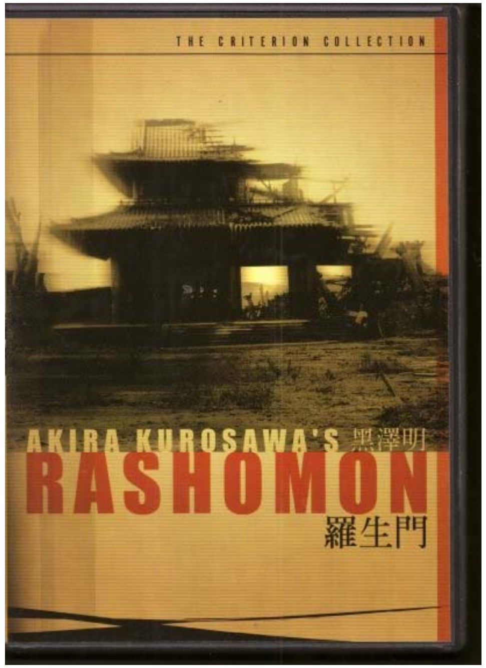
Leo Breiman, "Statistical Modeling: The Two Cultures",
Statistical Science, vol. 16, pp. 199-231, 2001.

… About the a*dvances of statistical modeling* …

Rashomon (羅生門,らしょうもん)

THE MULTIPLICITY OF GOOD MODELS





http://www.toshiromifune.org/images/lastsamurai/rashomon3.jpg

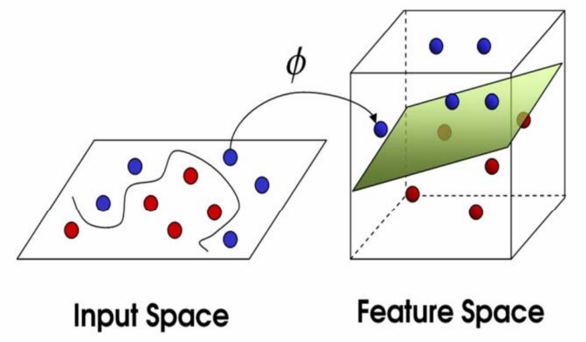
Occam

SIMPLICITY VS. ACCURACY

Figure adapted from wikimedia.org

Bellman

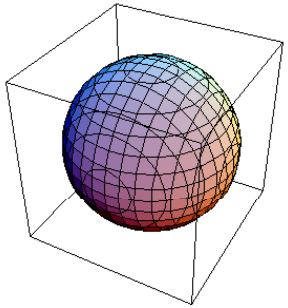
THE CURSE OF DIMENSIONALITY

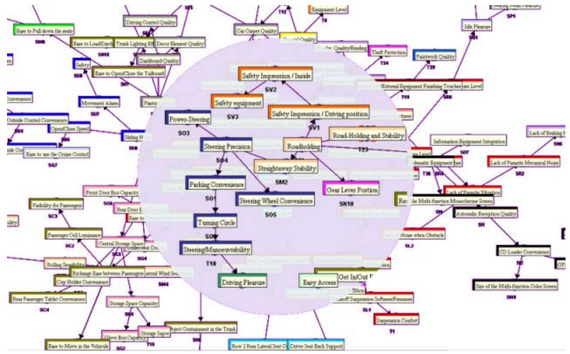


Pros:

- *Higher VC dimensions for linear classifiers*
- *Sparseness*

…





Cons:

- *Data in the tails*
- *Complexities of structural and parametric estimation*

…

Sword image from http://www.theswordepot.net/images/C-900S_SWORD.jpg
Hyperball figure from http://yaroslavvb.com/research/reports/curse-of-dim/pics/sphere.gif
Bayesian network figure from http://www.bayesia.com/assets/images/content/applications/en-analyse-questionnaires-satisfaction.jpg

The ultimate goal?

- Many new methods have been proposed
- New methods always look better than old methods
- More accurate solutions on more complicated problems

…

- Approach to the real computational intelligence ?

Photo from “A.I.”, Warner Bros.

Acknowledgement

- Promoters: Prof. B. De Moor and Prof. Y. Moreau
- Prof. J. A. K. Suykens
- Professors in the examination committee
- SISTA members: Tijl, Steven, Frizo, Bert, Carlos, Kristiaan, Sonia, Leo, Xinhai, Tunde, Fabian, Tillmann, Anneleen, Olivier
- Administrative staff: Ida
- Prof. M. Van Hulle (Neruofysiologie, Gasthuisberg)
- Prof. W. Glänzel (Steunpunt O & O, K.U.Leuven)
- Dr. J. Ye (Arizona State University)

Thank you!