BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 8(3)

Evaluation of data integration strategies based on kernel method of clinical and microarray data

Ary Noviyanto * & Ito Wasito

Faculty of Computer Science, Universitas Indonesia; Ary Noviyanto – E mail: ary.noviyanto@ui.ac.id; * Corresponding author

Received January 19, 2012; Accepted January 20, 2012; Published February 03, 2012

Abstract:
The cancer classification problem is one of the most challenging problems in bioinformatics. The data provided by Netherland Cancer Institute consists of 295 breast cancer patient; 101 patients are with distant metastases and 194 patients are without distant metastases. Combination of features sets based on kernel method to classify the patient who are with or without distant metastases will be investigated. The single data set will be compared with three data integration strategies and also weighted data integration strategies based on kernel method. Least Square Support Vector Machine (LS-SVM) is chosen as the classifier because it can handle very high dimensional features, for instance, microarray data. The experiment result shows that the performance of weighted late integration and the using of only microarray data are almost similar. The data integration strategy is not always better than using single data set in this case. The performance of classification absolutely depends on the features that are used to represent the object.

Background:
In bioinformatics data, an object can be represented by several heterogeneous data sets. The combination of heterogeneous data set based on kernel method is considered to produce better classification result **[1, 2, and 3].** One of the bioinformatics data that contains several heterogeneous data sets is data of breast cancer patients which consists of microarray data and clinical data **[4].** The objective is to classify the breast cancer patients that are with the distant cancer or without the distant cancer. Simply, distant cancer, it can be called as distant metastasis, indicates that the cancer spread from the primary tumor to distant organs or distant lymph nodes **[5].** Support Vector Machine (SVM) which is introduced by Vapnik, is a powerful classifier that is often used bioinformatics application **[6].** The problem in SVM is finding the proper parameters to build the model. It takes time while the data set or the features set is a large size, such as microarray data. Least Square Support Vector Machine (LS-SVM) is claimed as a modified SVM that can faster in training phase **[7].** The main idea of SVM is finding the best hyper plane that can be separated the data into two classes; positive and negative. Using kernel method, mixed data can be transformed into higher dimension, which it can be separated linearly, implicitly using kernel function **[6, 8].**

The means which are to combine several heterogeneous data sets can be called as data integration strategies. The various features sets can be integrated in simple manner based on kernel method **[1].** This research is closely related with a research by Daemen A *et al.*, "a kernel-based integration of genome-wide data for clinical decision support." Daemen A *et al.* concluded that the accuracy of cancer prediction was increase if the multiple data sets were integrated **[3].** The main contribution of this paper is to evaluate the performance of classification result in the term of the distant cancer classification which implements the data integration strategies and not. The data sets refer to 295 breast cancer patients which are public domain provided by Netherland Cancer Institute. This data consist of 101 patients are with and 194 patients are without distant metastases. The data set is spitted into training set and validation set. The training set contains 148 data which consist of 47 with and 101 without distant metastases and the validation set contains 147 which consist of 54 with and 93 without distant metastases. The characteristic of the data are woman who were younger than 53-years old and the tumor which are smaller than 5cm **[4].**

Microarray technology is very important tool to monitor genome wide expression level of genes in an organism. The microarray data contains of 24.496 spots or features which can be selected into 70 features that are good-prognosis signature. The clinical data contains only 13 variables; Diameter of tumor (Numeric), T1_T2 (Binary; ≤2cm or >2cm), pN (3 classes; pN0, 1-3, or ≥4), Number of positive Lymph nodes (Numeric), Mastectomy (Binary; yes or no), Estrogen Receptor (Binary; positive or negative), Tumor grade (3 classes; poorly, intermediate or well differentiated), Age (Numeric), Chemotheraphy (Binary; yes or no), Hormonal therapy (Binary; yes or no), St. Gallen criteria (Binary; Chemotherapy or no chemotherapy), National Institutes of Health (NIH) consensus criteria (Binary; Chemotherapy or no chemotherapy), NIH risk (3 classes; low, intermediate or high) **[4].** In order to know visually about the distribution data sets, the Kernel Dimensionality Reduction (KDR) **[9]** reduces the features dimension of microarray data from 70 to 2 and 3 as showed in **(Figure 1 and 2);** and also feature dimension of clinical data from 13 to 2 and 3 as showed in **(Figure 3 and 4)**. The blue circle and the red circle represent the two classes. Based on the visualizations, the data are mixed.

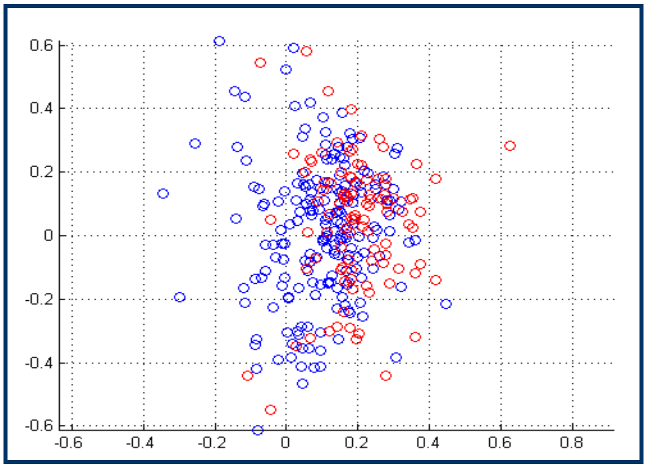

Figure 1: Visualization of reduced microarray data features into 2 dimensions.

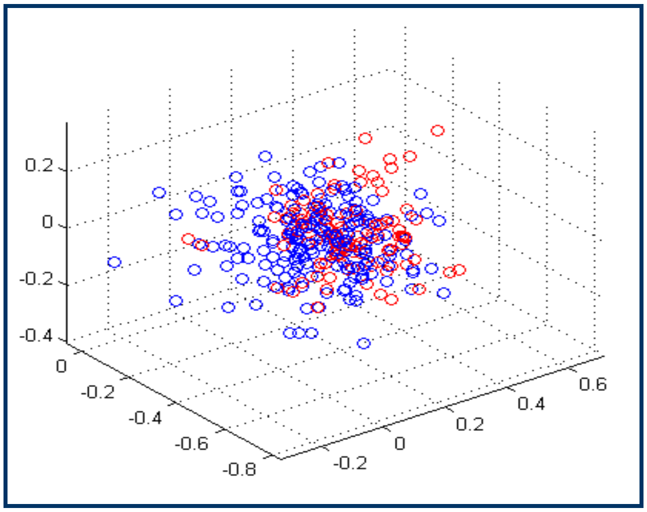

Figure 2: Visualization of reduced microarray data features into 3 dimensions.

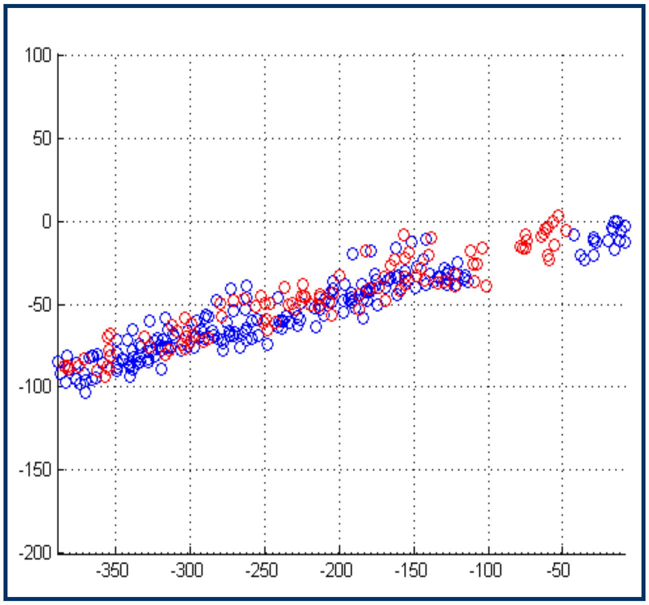

Figure 3: Visualization of reduced clinical data features into 2 dimensions.

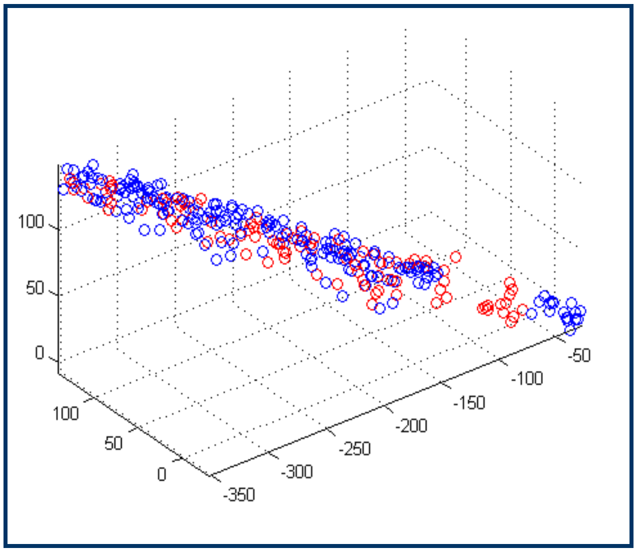

Figure 4: Visualization of reduced clinical data features into 3 dimensions.

Methodology:

The kernel based data integration strategies are classified in three classes; early integration, intermediate integration and late integration **[2].** The early integration is a concatenation of features sets, the intermediate integration is a combination of kernel matrices and the late integration is a combination of the models. Three common kernel functions will be used in this experiment; linear kernel, polynomial kernel and Radial Basis Function (RBF) kernel. The equations to combine the kernel matrices in the intermediate integration are formulated according to Equation (1) for linear kernel, Equation (2) for polynomial kernel and Equation (3) for RBF kernel in the Supplementary material. These kernel functions are applied to LS-SVM. The differences of LS-SVM and SVM is that LS-SVM uses square loss function instead of hinge loss function so that the model's parameters can be solved linearly instead of quadratic programming **[7].** The technique to get regularization parameter and kernel parameters is by using a combination of Coupled Simulated Annealing (CSA) and a standard simplex method with leave-one-out cross validation strategy **[10].**

The issue of different signification level of each feature sets leads to give the weight to each features sets. In this paper, it is called as weighted data integration. It means that each features sets will be weighted with a particular real number so that it controls the contribution of the features set to get the final conclusion. The implementation of weighted data integration strategy is straight forward by multiply kernel matrix with a real value. In this case of using two feature sets (i.e. micro-array data set and clinical data set), the total weighted of these two data set is 1.0. The equations of the weighted intermediate integration are formulated according to Equation (4) for linear kernel, Equation (5) for polynomial kernel and Equation (6) for RBF kernel in then the Supplementary material.

Discussion:

The result of the experiment scenario can be showed in **Table 1(see supplementary material).** The best performance in AUC, 0.7493, is weighted late integration using two RBF kernel functions which the configuration are weight of 0.74 for microarray data set, weight of 0.26 for clinical data set, gamma parameter of 3.317 for microarray data set, gamma parameter of 1.685 for clinical data set, σ^2 of 39.686 for microarray data set and σ^2 of 13.276 for clinical data set. The best kappa statistic, 0.4725, is model of only microarray data set using RBF kernel function which the configurations are 3.380 for gamma parameter and 46.918 for σ^2. These two models have similar accuracy, 0.7415, and also become the best accuracy. The weighted data integration strategy (i.e. weighted intermediate integration and weighted late integration) show better performance than if it is treated as same weight. In the single data set, clinical data shows worse performance than microarray data in every kernel functions. The AUC of clinical data is no more than 0.6605 and the AUC of microarray data can reach 0.7421. The performance of kernel function is moderate. It really depends on the value of kernel parameter that is used. Overall the best performance kernel function is RBF.

Conclusion:

The experiment scenario is designed to evaluate the performance of data integration strategy and the using of single data set in the case of breast cancer classification. The issue of significant level of features set is also included as weighted data integration strategy. The experiment scenario contains 33 models that are evaluated. The complete result can be shown in **Table 1 (see supplementary material)** The data is taken from public domain from Nederland Cancer Institute that contains microarray data set and clinical data set as supplementary data.

Based on the experiments result, the microarray his little bit higher value of kappa statistic than the weighted late but the value of AUC of the weighted late has little bit higher than the AUC of the microarray. It shows that the performance of data integration strategy is almost similar with only using single data set. For further analysis it can be inferred that the clinical data set shows relatively bad performance. The experiment using microarray produce similar performance with the weighted late because the clinical data set in the weighted late cannot contribute much to the classification result. Generally, the classification performance is controlled by the features sets that are used.

References:

[1] Lanckriet GR *et al. Bioinformatics*. 2004 **20:** 2626 [PMID: 15130933].

[2] Ozen A BMC *et al. Struct Biol*. 2009 **9**: 66 [PMID: 19840377]

[3] Daemen A *et al. Genome Med*. 2009 **1**: 39 [PMID: 19356222]

[4] Van de Vijver MJ *et al. N Engl J Med*. 2002 **347**: 1999 [PMID: 12490681].

[5] http://www.nccn.com/component/glossary/Glossary-1/D/page,2/

[6] Ben-Hur A *et al. PLoS Comput Biol*. 2008 **4**: e1000173 [PMID: 18974822]

[7] Suykens JAK & Vandewalle J, Neural Process. Lett. 1999 **9**: 3

[8] Sánchez A VD. *Neurocomputing*. 2003 **55**:

[9] Fukumizu K *et al. The Annals of Statistics*. 2009 **37**: 4

[10] http://www.esat.kuleuven.be/sista/lssvmlab/

Edited by P Kangueane
Citation: Noviyanto & Wasito, Bioinformation 8(3): 147-150 (2012)

Supplementary material:

The experiments will evaluate the performance of the data integration strategies compared with only using single features using various kernel functions. The experiment also compares the unweighted and weighted of data integration. The best performance weights are searched using brute force approach which the range of weight is from 0.1 to 0.99 and increment of 0.01. If the weight of microarray data set is *a* then the weight of clinical data set is 1-*a*. Accuracy, AUC and Kappa statistic are used as the performance evaluation. Accuracy show the correct classification rate of the classification process given certain constant threshold whereas AUC is more general form of accuracy which represents ROC curve in single value. The AUC is an area under ROC (Receiver Operating Characteristics) curve with range value is between 0.0 and 1.0. The ROC is a two dimension graph that plots (1-specitifity) and sensitivity. The Kappa statistic shows the agreement level between the classification result and the ground truth. It means that how match the classification model with the real model.

$K = X_1 \cdot X_1^T + X_2 \cdot X_2^T$	Where X_1 is the matrix of first features set, X_2 is the matrix of second features set and K is the combination kernel. The X matrix has row as the number of data and column as the number of features.	Integration using linear kernel → (1)
$K = (X_1 \cdot X_1^T + t)^d + (X_2 \cdot X_2^T + t)^d$	Where t (intercept) and d (degree of the polynomial) are the kernel parameter.	Integration using polynomial kernel → (2)
$K = \exp\left[-\left(\lVert X_1\rVert^2 * one + (\lVert X_1\rVert^2 * one)^T - 2X_1 \cdot X_1{}^T\right)/(2*\sigma^2)\right]$ $+ \exp\left[-\left(\lVert X_2\rVert^2 * one + (\lVert X_2\rVert^2 * one)^T - 2X_2 \cdot X_2{}^T\right)/(2*\sigma^2)\right]$	Where σ^2 is the variance of the kernel, *one* is a matrix with the value of each element is 1 and the size is *n*x*n*; *n* is the number of data. $\lvert\lvert X \rvert\rvert^2$ means square of the norm value of each row or each data.	Integration using RBF kernel→ (3)
$K = w_1(X_1 \cdot X_1^T) + w_2(X_2 \cdot X_2^T)$	Where w_1 is a weight for data set 1 and w_2 is a weight for data set 2. w_1 = 1- w_2.	Weighted integration using linear kernel→ (4)
$K = w_1(X_1 \cdot X_1^T + t)^d + w_2(X_2 \cdot X_2^T + t)^d$		Weighted Integration using polynomial kernel → (5)
$K = w_1\left(\exp\left[-\left(\lVert X_1\rVert^2 * one + (\lVert X_1\rVert^2 * one)^T - 2X_1 \cdot X_1{}^T\right)/(2*\sigma^2)\right]\right)$ $+ w_2\left(\exp\left[-\left(\lVert X_2\rVert^2 * one + (\lVert X_2\rVert^2 * one)^T - 2X_2 \cdot X_2{}^T\right)/(2*\sigma^2)\right]\right)$		Weighted Integration using RBF kernel→ (6)

Table 1: Experimental Result

No	Data set or Data integration strategies	Kernel function		W_1	W_2	gamma		Kernel parameters				Acc	AUC	std	Kappa
								σ^2 or t		d					
1	Clinical	Linear		-	-	0.007		-		-		0.6599	0.6605	0.0449	0.1717
2		Poly		-	-	0.127		0.609		5		0.6327	0.4998	0.0505	0.0097
3		RBF		-	-	59.944		251.015		-		0.6599	0.6454	0.0464	0.2956
4	Microarray	Linear		-	-	0.021		-		-		0.7075	0.7031	0.0463	0.3633
5		Poly		-	-	592.288		5.117		5		0.6667	0.6639	0.0454	0.3223
6		RBF		-	-	3.380		46.918		-		0.7415	0.7421	0.0426	0.4725
7	Early	Linear		-	-	0.011		-		-		0.7075	0.7065	0.0449	0.3318
8	Integration	Poly		-	-	3.264		5.628		4		0.6667	0.6685	0.0455	0.3018
9		RBF		-	-	9414.010		29.314		-		0.6803	0.7109	0.0426	0.3354
10	Intermediate	Linear		1.00	1.00	1.566		-		-		0.6871	0.6509	0.0472	0.2576
11	Integration	Poly		1.00	1.00	0.093		46.110		3		0.6939	0.6838	0.0457	0.3230
12		RBF		1.00	1.00	6922.329		41.927		-		0.6735	0.6912	0.0441	0.3481
13	Weighted	Linear		0.55	0.45	0.012		-		-		0.7143	0.7172	0.0442	0.3391
14	Intermediate	Poly		0.35	0.65	0.059		34.159		4		0.7143	0.6942	0.0449	0.4127
15	Integration	RBF		0.90	0.10	2.420		31.842		-		0.7211	0.7459	0.0416	0.4412
16	late	Linear	Linear	1.00	1.00	0.023	0.006	-	-	-	-	0.7279	0.7117	0.0451	0.4235
17	Integration	Linear	Poly	1.00	1.00	0.025	0.127	-	0.609	-	5	0.6327	0.5546	0.0502	0.0097
18		Linear	RBF	1.00	1.00	0.031	1.406	-	11.459	-	-	0.7211	0.6943	0.0459	0.3680
19		Poly	Linear	1.00	1.00	592.288	0.006	5.117	-	5	-	0.6599	0.6724	0.0452	0.3400
20		Poly	Poly	1.00	1.00	592.288	0.127	5.117	0.609	5	5	0.6327	0.5139	0.0502	0.0097
21		Poly	RBF	1.00	1.00	592.288	274.241	5.117	1218.462	5	-	0.6531	0.6573	0.0458	0.2622
22		RBF	Linear	1.00	1.00	76.353	0.006	53.602	-	-	-	0.7143	0.7222	0.0438	0.4038
23		RBF	Poly	1.00	1.00	27.094	0.127	64.577	0.609	-	5	0.6327	0.5671	0.0496	0.0097
24		RBF	RBF	1.00	1.00	17.166	33.055	68.646	429.924	-	-	0.7007	0.7163	0.0436	0.3937
25	Weighted late	Linear	Linear	0.57	0.43	0.022	0.007	-	-	-	-	0.7279	0.7135	0.0452	0.4235
26	Integration	Linear	Poly	0.99	0.01	0.031	0.127	-	0.609	-	5	0.7007	0.6904	0.0467	0.3510
27		Linear	RBF	0.40	0.60	0.023	104.368	-	16840.814	-	-	0.7075	0.7117	0.0446	0.3582
28		Poly	Linear	0.46	0.54	592.288	0.006	5.117	-	5	-	0.6667	0.6740	0.0451	0.3599
29		Poly	Poly	0.99	0.01	592.288	0.127	5.117	0.609	5	5	0.6463	0.6454	0.0458	0.2938
30		Poly	RBF	0.62	0.38	592.288	50.596	5.117	8173.148	5	-	0.6803	0.6736	0.0451	0.3817
31		RBF	Linear	0.62	0.38	3.079	0.153	45.272	-	-	-	0.7279	0.7457	0.0417	0.4448
32		RBF	Poly	0.97	0.03	2.996	0.127	38.101	0.609	-	5	0.7279	0.7272	0.0427	0.4322
33		RBF	RBF	0.74	0.26	3.317	1.685	39.686	13.276	-	-	0.7415	0.7493	0.0420	0.4565