Supplementary material for Manuscript BIOINF-2005-1602

Title: Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data

Appendix A. Testing K-Nearest Neighbor and Support Vector Machine

To further evaluate the proposed oMLP, we compare the oMLP with k-nearest neighbor (KNN) and a type of multi-class support vector machine (SVM), One-vs-Rest SVM (OVR-SVM), on the same three microarray data sets already tested by the oMLP and cMLP. As shown in comparison studies, the performance of OVR-SVM is typical in the multi-class SVM group [Statnikov *et al.*, 2005].

K-Nearest Neighbor

The Algorithm of KNN

Find the *K* samples in the training set that are closest to the unknown sample, and then predict the class of the unknown by majority vote, *i.e.*, choose the class that is most common among those *K* closest samples. We used Euclidean distance as the distance measure. The parameter *K* was determined by 100 iterations of 3-fold cross validation (all results are presented in Appendix A1).

Inputs of KNN

The KNN took optimal JDG set as inputs.

One-vs-Rest Support Vector Machine

The Structure of OVR-SVM

The OVR-SVM built with *K* binary SVMs can perform a *K*-class classification task; each binary SVM unit is trained to separate a specific class from the rest (Ramaswamy *et al.*, 2001; Statnikov *et al.*, 2005) (dash lines in Figure 1). A test sample is assigned to a class whose associated binary SVM has the largest real-valued output (solid lines in Figure 1).

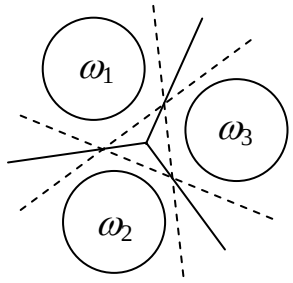


Figure 1. The structure of an OVR-SVM.

The Training of Binary Support Vector Machines

Each binary SVM in the OVR-SVM has the same kernel function and parameters. Given a training set

$X = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subset R^d \times \{-1, 1\}$,

the solution of a binary SVM is derived by a quadratic programming problem [Vapnik, 1998],

$$\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) + C \cdot \sum_{i=1}^{n}\xi_i ,$$

$$\text{s.t. } y_i(\sum_{j=1}^{n}\alpha_j y_j k(\mathbf{x}_i, \mathbf{x}_j) + b) \geq 1 - \xi_i,\ i = 1, \ldots, n,$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n.$$

We choose the penalty constant *C* from a set {0.001, 0.01, 0.1, 1.0, 10.0} to achieve the lowest cross validation error. The derived decision function is,

$$f(\mathbf{x}) = \mathrm{sgn}(\sum_{i \in S} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b),$$

where *S* is the index set of the support vectors and $k(\mathbf{x}, \mathbf{z})$ denotes the kernel function.

Kernel Functions

We tested 3 kernel functions (Vapnik, 1998) with different kernel parameters in this study:

Linear kernel

$$k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x} \cdot \mathbf{z} \rangle,$$

where $\langle \mathbf{x} \cdot \mathbf{z} \rangle$ is the inner product of vectors $\mathbf{x}$ and $\mathbf{z}$.

Polynomial kernel

$$k(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x} \cdot \mathbf{z} \rangle + 1)^q,$$

where q is the degree of the polynomial kernel. We tested two polynomial kernels with $q = 2$ and $q = 3$.

Gaussian kernel

$$k(\mathbf{x}, \mathbf{z}) = \exp(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{\sigma \cdot d}),$$

where d is the dimension of the sample space and σ is the scaling factor that makes the width of the Gaussian function be proportional to d. We tested four Gaussian kernels with scaling factors, $\sigma = 0.01$, $\sigma = 0.1$, $\sigma = 0.5$, and $\sigma = 1.0$.

Table 1. The list of abbreviations of kernel functions.

Abbreviation	Kernel
Linear	Linear kernel
Poly 2	Polynomial kernel, degree 2
Poly 3	Polynomial kernel, degree 3
Gaus 0.01	Gaussian kernel, scaling factor 0.01
Gaus 0.1	Gaussian kernel, scaling factor 0.1
Gaus 0.5	Gaussian kernel, scaling factor 0.5
Gaus 1.0	Gaussian kernel, scaling factor 1.0

Table 1 gives a list of abbreviations of all tested kernel functions in this study.

Inputs of OVR-SVM

The OVR-SVMs took two types of inputs: optimal JDG set and all genes.

Results

Table 2 summarizes the performances of oMLP, cMLP, KNN and OVR-SVM. For the KNN and OVR-SVM, we listed the model with the best performance in Table 2, and presented the results of all tested models of KNN and OVR-SVM in Appendix A1. Note that average and STD of classification rate are calculated based on 100 iterations of 3-fold cross validations.

Table 2. The summary of performances of MLP, KNN and SVM. The classification rate (%) is listed as average (STD).

				OVR-SVM	
Data set	**oMLP**	**cMLP**	**KNN**	**Optimal JDG**	**All genes**
LGMD	98.69 (4.39)	42.05 (18.96)	41.33 (12.66) K = 15	100.00 (0.00) Linear, *C* = 0.01, 0.1, 1.0, 10.0 Gaus 10.0, *C* = 10.0	50.94 (12.58) Gaus 10.0, *C* = 10.0
Leukemia	96.96 (5.27)	87.37 (15.7)	88.39 (8.73) K=6	98.37 (3.76) Linear, *C* = 10.0	95.34 (5.97) Linear, *C* = 1.0
CNS cancer	89.82 (4.46)	86.86 (7.19)	86.59 (4.65) K=4	95.59 (3.25) Gaus 10.0, *C* = 10.0	89.13 (3.49) Linear, *C* = 1.0

In summary, the OVR-SVM with optimal JDG set as inputs and the oMLP had comparable performances, and the OVR-SVM demonstrated small improvements, whereas the KNN and cMLP showed much poorer performances in all cases.

Furthermore, we compared the oMLP and *binary* SVM on 10 two-class microarray data sets that are subsets of multi-class sets (Table 3). Both oMLP and SVM took optimal JDG subset as inputs. Again, the performances of the oMLP and SVM are comparable in all cases (Table 4).

Table 3. The list of microarray data sets tested by the oMLP and binary SVM.

Dataset	Number of Classes	Number of Samples	Source	Name of Classes	Number of Genes
Leukemia_Golub	2	72	(Golub *et al.* 1999)	ALL (n=47), AML (n=25)	7,129
SRBCT	2	43	(Khan J. *et al.* 2001)	EWS (n=23), RMS (n=20)	2,308
CNS_A2_1	2	70	(Pomeroy *et al.* 2002)	Brain_MD (n=60), Brain_Rhab (n=10)	7,129
MD_FSH_NHM	2	32	CNMC (MAS 5.0 A chip)	FSH (n=14), NHM (n=18)	11,252
MD_Calpain3_Dysferlin	2	20	CNMC (dCHIP B chip)	Calpain3 (n=10), Dysferlin (n=10)	8,466
MD_supergroup	2	82	CNMC (MAS 5.0 A chip)	Class 1: BMD (n=5), Calpain3 (n=10), DMD (n=10), Dysferlin (n=10), FKRP (n=7); Class 2: EDMD1 (n=4), EDMD2 (n=4), FSH (n=14), NHM (n=18)	11,252
Lung_cancer_A	2	156	(Bhattacharjee *et al.* 2001)	AD (n=139), NL (n=17)	12,600
Breast_cancer	2	78	('t Veer *et al.* 2002) (log10 intensity)	metastatic disease (n=34), disease free (n=44)	24,481
CNS_B	2	34	(Pomeroy *et al.* 2002)	Brain_MD subtype classic (n=25), Brain_MD subtype desmoplastic (n=9)	7,129
CNS_A2_2	2	70	(Pomeroy *et al.* 2002)	Brain_MD (n=60), Brain_MGlio (n=10)	7,129

Table 4. The summary of prediction accuracies of the oMLP and binary SVM. For each data set, the best model(s) among all test SVM models are bolded.

			Binary SVM						
				Polynomial Kernal		Gaussian Kernel			
Dataset	Size of Optimal JDG subset	oMLP	Linear Kernel	$q = 2$	$q = 3$	$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$
Leukemia_Golub	9	100.00%	**100%**	99.97%	98.64%	65.29%	68.79%	99.99%	**100%**
SRBCT	10	100%	**100%**	99.91%	99.39%	53.50%	85.20%	99.33%	**100%**
CNS_A2	4	99.30%	**99.89%**	99.83%	99.60%	85.75%	85.75%	99.56%	**99.89%**
MD_FSH_NHM	16	99.51%	**99.91%**	95.82%	96.92%	56.36%	83.42%	96.49%	98.06%
MD_Calpain3_Dysferlin	12	100%	**100%**	98.43%	99.25%	66.31%	62.64%	98.47%	**100%**
MD_supergroup	19	96.99%	**97.84%**	93.60%	93.07%	51.24%	84.20%	93.86%	94.85%
Lung_cancer_A	18	98.59%	98.26%	97.72%	97.43%	89.11%	89.11%	97.66%	**98.32%**

Breast_cancer	9	96.83%	**98.35%**	92.92%	94.53%	56.41%	78.13%	95.60%	96.58%
CNS_B	5	98.09%	98.61%	95.89%	91.62%	73.48%	73.49%	94.43%	**99.24%**
CNS_A2	29	99.79%	**99.97%**	97.88%	97.17%	85.75%	85.75%	88.86%	99.59%

Appendix B. Testing untrained optimized MLP and conventional MLP

As described in the manuscript, if initialized using the proposed MLP initialization method, the hidden layer of an *untrained* oMLP is able to extract discriminant features derived from wFC; the neurons in the output layer can perform linear one-vs-rest classifications based on these extracted features. We used *linear* transfer function in the hidden neurons and *log-sigmoid* transfer function in the output neurons in the experiments presented in the manuscript. Hence, an *untrained* oMLP closely resembles LDA, and the initial condition of the oMLP (*i.e.*, performance of the untrained oMLP) reflects the performance of LDA.

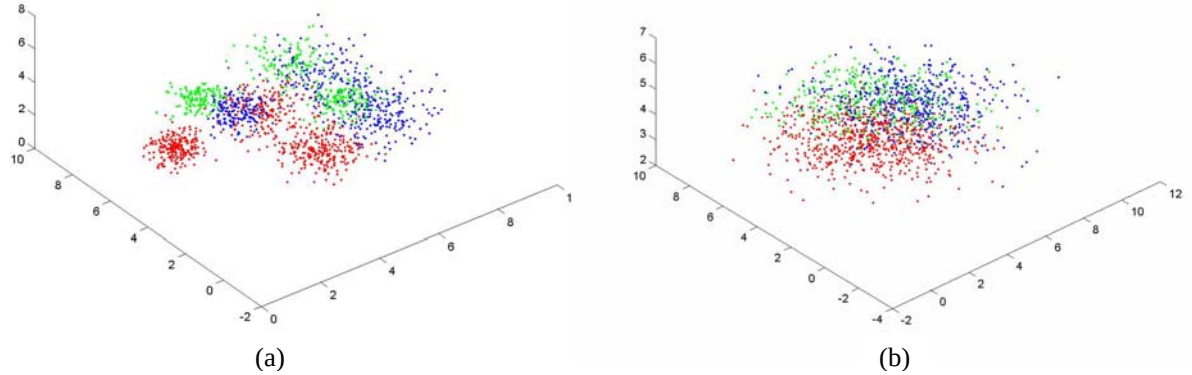
To further assess the effect of linear and nonlinear transfer functions in hidden neurons, we compared the prediction accuracies of *untrained* oMLPs and cMLPs with hidden neurons having linear or log-sigmoid functions (Table 5). The untrained oMLP with linear transfer function in hidden neurons generally had better classification rate than the untrained oMLP with log-sigmoid transfer function in each tested case. Besides, when hidden neurons had *linear* transfer function, the untrained oMLPs considerably outperformed the untrained cMLPs; on the other hand, the difference between the untrained oMLPs and cMLPs becomes negligible with *log-sigmoid* function in hidden neurons.

Table 5. The performances of untrained oMLP and cMLP.

Data set	**Transfer function in hidden neurons**	**Classifier**	**Classification rate, %**	
			Average	**STD**
LGMD	Linear	oMLP	80.07	13.36
		cMLP	21.57	13.74
	Log-sigmoid	oMLP	27.60	12.84
		cMLP	24.28	11.20
Leukemia	Linear	oMLP	38.35	11.40
		cMLP	31.47	17.11
	Log-sigmoid	oMLP	33.81	9.24
		cMLP	32.37	11.05
CNS cancer	Linear	oMLP	41.79	30.12
		cMLP	22.04	17.90
	Log-sigmoid	oMLP	28.33	27.17
		cMLP	22.72	24.97

Appendix C. Testing the limitation of the proposed MLP initialization method

We designed the following experiment to investigate whether the effectiveness of the proposed initialization method might diminish while the distribution of an individual class deviated from a single standard Gaussian to a mixed Gaussian. We simulated two types of three-dimensional three-class data sets: complex-Gaussian mixture (CGM) with each class having a mixed Gaussian distribution, and simple-Gaussian mixture (SGM) with each class having a single Gaussian distribution. For every CGM set, there is a corresponding SGM set in which each class is generated based on the structural parameters (class center and covariance) of a corresponding class in this CGM set (Figure 2). In other words, the corresponding classes in the CGM and SGM have the same structural parameters, but different distributions. Therefore, the MLPs for classifying the pair of CGM and SGM sets will have the same initial values that are calculated using the structural parameters. By testing the oMLP and cMLP on the pairs of CGM and SGM sets, we may observe whether and how the proposed method may be affected when the distribution of an individual class departs from a single Gaussian. Furthermore, we also adjusted the distances between the sub-clusters in each class in the CGM sets to make the distribution of a class more or less apart from a single Gaussian. Two distance parameters, C_1 and C_2, are used to control the distances between the classes and the distances between the sub-clusters in each class respectively. With larger C_1, individual classes are dragged far apart so that they have less overlaps, and subsequently higher classification accuracy will be expected. With larger C_2, the sub-clusters within each class are more distant from each other so that the distribution in each class is more deviate from the single Gaussian.



(a) (b)

Figure 2. Plots of simulated data sets, (a) CGM2 set (C_1 = 4, C_2 = 3), (b) SGM2 set.

We assessed the capability of the proposed initialization method by testing both oMLP and cMLP on each pair of CGM and SGM sets, and used two measures to evaluate the improvement made by oMLP over cMLP: percentage of increase in average of classification rate, and percentage of reduction in STD of classification rate (Table 6). Note that all results were calculated based on 100 iterations of 3-fold cross validations. Based on the results, we made two types of comparisons as follows.

1. Comparing the improvements by oMLP over cMLP for each CGM set and its corresponding SGM set may help find out whether the effectiveness of the wFC-based initialization is affected by the distortion of Gaussian distribution in each class. For example, the improvements in CGM1 set and SGM1 set are: 5.01% vs. 6.44% increase in average and 84.85% vs. 83.72% reduction in STD.
2. Compare the improvement by oMLP over cMLP

3. The slight differences exiting in all paired CGM and SGM sets indicated that

The similar percentages for paired CGM and SGM sets indicated that the effectiveness of the wFC-based initialization was not much affected by the distortion of Gaussian distribution in each class. Furthermore, the small decrease in these percentages when C_2

The experimental results showed the proposed MLP initialization method was not sensitive to the distortion of Gaussian distribution in each class.

Table 6. The comparison of the improvements made by oMLPs when classifying CGM and SGM sets. The improvement measures include the percentage of increase in the average of classification rate and percentage of reduction in the STD of classification rate.

		Classification rate			
Data set	**Classifier**	**Average**	**Increase of average**	**STD**	**Reduction of STD**
CGM1 set,	oMLP	82.68%	5.01%	1.35%	84.85%
$C_1 = 4$, $C_2 = 2$	cMLP	78.74%		8.90%	
SGM1 set,	oMLP	80.60%	6.44%	1.51%	83.72%
(corresponding to CGM_1)	cMLP	75.73%		9.29%	
CGM2 set,	oMLP	76.78%	3.85%	1.49%	80.10%
$C_1 = 4$, $C_2 = 3$	cMLP	73.93%		7.51%	
SGM2 set,	oMLP	71.45%	2.03%	1.67%	67.75%
(corresponding to CGM_2)	cMLP	70.02%		5.16%	
CGM3 set,	oMLP	80.60%	6.44%	1.51%	83.72%
$C_1 = 7$, $C_2 = 2$	cMLP	75.73%		9.29%	
SGM3 set,	oMLP	95.90%	4.84%	0.81%	92.72%
(corresponding to CGM_3)	cMLP	91.47%		11.19%	
CGM4 set,	oMLP	94.36%	5.98%	1.24%	88.55%
$C_1 = 7$, $C_2 = 3$	cMLP	89.04%		10.78%	
SGM4 set,	oMLP	91.21%	5.06%	0.97%	90.69%
(corresponding to CGM_4)	cMLP	86.82%		10.42%	

References:

Bhattacharjee, Arindam *et al.* (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad Sci*, **98**(24), 13790-13795.