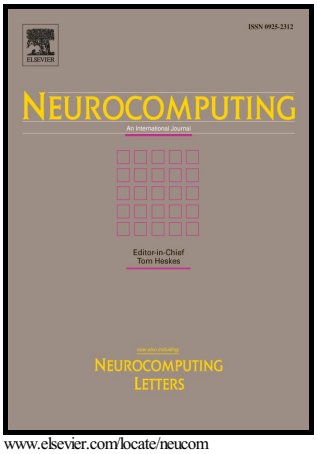

www.elsevier.com/locate/neucom

Local Partial Least Square Classifier in High Dimensionality Classification

Weiran Song, Hui Wang, Paul Maguire, Omar Nibouche

PII: S0925-2312(16)31581-8
DOI: http://dx.doi.org/10.1016/j.neucom.2016.12.053
Reference: NEUCOM17884

To appear in: *Neurocomputing*

Received date: 23 February 2016
Revised date: 10 June 2016
Accepted date: 18 December 2016

Cite this article as: Weiran Song, Hui Wang, Paul Maguire and Omar Nibouche, Local Partial Least Square Classifier in High Dimensionality Classificatior *Neurocomputing,* http://dx.doi.org/10.1016/j.neucom.2016.12.053

Local Partial Least Square Classifier in High Dimensionality Classification

Weiran Song, Hui Wang, Paul Maguire, Omar Nibouche.
Affiliation: Ulster University

Contact information.
Weiran Song: song-w@email.ulster.ac.uk
Hui Wang: h.wang@ulster.ac.uk
Paul Maguire: pd.maguire@ulster.ac.uk
Omar Nibouche: o.nibouche@ulster.ac.uk

Abstract

A central idea in distance-based machine learning algorithms such k-nearest neighbors and manifold learning is to choose a set of references, or a neighborhood, based on a distance functions to represent the local structure around a query point and use the local structures as the basis to construct models. Local Partial Least Square (local PLS), which is the result of applying this neighborhood based idea in Partial Least Square (PLS), has been shown to perform very well on the regression of small-sample sized and multicollinearity data, but seldom used in high-dimensionality classification. Furthermore the difference between PLS and local PLS with respect to their optimal intrinsic dimensions is unclear. In this paper we combine local PLS with non-Euclidean distance in order to find out which measures are better suited for high dimensionality classification. Experimental results obtained on 8 UCI and spectroscopy datasets show that the Euclidean distance is not a good distance function for use in local PLS classification, especially in high dimensionality cases; instead Manhattan distance and fractional distance are preferred. Experimental results further show that the optimal intrinsic dimension of local PLS is smaller than that of the standard PLS.

Keywords: High Dimensionality Classification, Distance function, Fractional Distance, Local Partial Least Squares.

1. Introduction

High-dimensional data are quite common in the real world. Such data are usually contaminated by outliers, irrelevant or redundant dimensions, which could invalidate data mining algorithms, yielding biased and inaccurate models. To circumvent this issue, noise reduction [1] and dimensionality reduction [2, 3] are usually applied in order to obtain suitable and effective representation of data. Partial Least Square (PLS) is initially a latent modeling approach for linear regression [4] and later on extended to PLS discriminant analysis (PLS-DA) for classification [5], which is able to analysis the small sample size and multicollinearity data such as chemical spectroscopy data. These data are usually high-dimensional and highly correlated in neighboring independent variables. Such type of data will invalidate Fisher discriminant analysis (FDA) and local FDA as scatter matrices become singular [6, 7]. Basically, PLS handles the high-dimensional problem by means of latent projection before modelling in order to avoid the effects of the curse of dimensionality [8].

In recent studies, PLS has been effectively extended to local PLS algorithm to handle data with complex global structures and data with characteristics such as nonlinearity and multimodal distributions using local learning approach [9, 10]. This approach uses neighboring points to predict a query point by modelling in local space [11], thus any global influence can be lessened. For example, locally weighted PLS (LW-PLS), which utilizes the similarity based on weighted Euclidean distance, is proposed to estimate the content of active pharmaceutical ingredients and drastically improved by 38.6% in root mean square error of prediction (RMSEP) compared with PLS [9]. Similarly, local regression (LR) approach is coupled with PLS to predict the soil organic carbon content [12] and analyze the voice conversion [13]. Some modified algorithms aim to further improve the performance of local PLS methods by adapting the distance functions used. A covariance-based distance scheme is applied in [14] which explicitly takes account of the relationship among variables and reduces the computational load. Another study utilizes cosine distance after locality preserving projection (LPP) [15] to measure the sample similarities and establishes a high-performance calibration model [16]. In addition, a non-Euclidean function defined in [17] has been used for the detection and measurement of residual drug substances; the results of using such a distance exceeded those of a Euclidean norm-based by over 10% [18]. As a result, the performances of distance functions selected in local PLS can be varying and Euclidean distance may not always be the best option.

In fact, it has been proven that Euclidean distance can barely make sufficient distinctions between different data points in high-dimensional case [19]. Given a query point, the ratio of its Euclidean distances with its nearest and farthest neighbors approaches 1 under broad conditions as dimensionality increases [20]. In such case, Manhattan or fractional distance are preferred and can provide better discriminations [19, 20]. As PLS-DA is commonly used to analyze high-dimensional data in literatures, it is therefore reasonable to expect better classification performance if fractional distance is embedded in local PLS-DA.

In this paper, we combine local PLS-DA with non-Euclidean distance, specifically fractional distance to classify data with varying dimensionalities. Furthermore, as dimensionality reduction is embedded in PLS, the optimal intrinsic dimension is related to latent variables. If a complex global structure is simplified in local PLS, a reduction in the number of latent variables can be expected. In this paper, we also study the optimal intrinsic dimensions in both local and global (ordinary) PLS. The results show that non-Euclidean distance, such as fractional and cosine are preferable than Euclidean distance in local PLS-DA. Meanwhile, the number of latent variables (LVs) in local PLS-DA is smaller than that in the global PLS.

2. Local PLS for classification

2.1 PLS regression

PLS problem is to establish a relationship between independent variables X_{nd} and dependent variables Y_{nq} in latent space towards the regression model:

$$\min \| Y_{nq} - X_{nd} B \| \qquad (1)$$

where B is a d-by-q regression coefficient matrix. X and Y are decomposed as follows:

$$X = TP' + E$$

$$Y = UQ' + F \quad (2)$$

(3)

Here, T and U are low-dimensional latent representations of X and Y with size of $n \times r$. P and Q are termed loading matrices with size of $p \times r$ and $q \times r$. E and F are residual matrices. Suppose there is a linear relationship such that:

$$U = TD + H \quad (4)$$

where D is an r-by-r diagonal matrix, then

$$Y = TC' + F^* \quad (5)$$

where

$$C' = DQ' \quad (6)$$

and

$$F^* = HQ' + F \quad (7)$$

Based on (2) and (5),

$$Y = XP\left(P^T P\right)^{-1} C^T + F^* - EP\left(P^T P\right)^{-1} C^T \quad (8)$$

The PLS regression coefficient

$$B_{PLS} = P\left(P^T P\right)^{-1} C' \quad (9)$$

A number of algorithms can be used for solving the PLS regression problem. Among such algorithms, the Simple Partial Least Squares (SIMPLS) is a non-iterative algorithm that provides a fast and efficient computation [21]. In SIMPLS, X and Y are standardized as X_0 and Y_0, respectively. Then Singular Value Decomposition (SVD) are applied on ${X_0}^T Y_0$ and its deflations to maximize covariance for multivariate Y . In this paper, the PLS based experiments are operated by SIMPLS algorithm.

Algorithm: SIMPLS

Compute cross-product: $S = {X_0}^T Y_0$

for i = 1 to LVs

if i = 1 compute SVD of S

if i > 1 compute SVD of $S - P(P^T P)^{-1} P^T S$

get weights: s = first left singular vector

compute scores $t = X_0 s$

compute loadings $p = X_0^T t / (t^T t)$

store s, t, and p into R, T, and P respectively

end

Compute regression coefficients $B_{PLS} = R\left(T^T T\right)^{-1} T^T Y_0$

2.2 PLS-DA

Given that PLS is closely related to canonical correlation analysis (CCA) and CCA is related to linear discriminant analysis (LDA) in turn, reference [5] transforms PLS regression to classification by using dummy matrix to represent category information. The category information is initiated as a null matrix Y in which rows and columns are equal to the number of instances n and categories *c*, respectively. If the values of a category c_i integrally increase from one, the corresponding element in c_i-th column of Y is set to 1. Suppose 1_k is a $k \times 1$ vector of all ones and, 0_k is a $k \times 1$ vector of all zeros. After arranging the sequence of category information, Y can be reordered as:

$$Y = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_c} & 0_{n_c} & \cdots & 1_{n_c} \end{pmatrix}_{n \times c} \quad (10)$$

or

$$Z = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & 1_{n_c-1} \\ 0_{n_c} & 0_{n_c} & \cdots & 0_{n_c} \end{pmatrix}_{n \times (c-1)} \quad (11)$$

where $\sum_{j=1}^{c} n_j = n$. Then the category $Y_{predict}$ (*1*-by-*c*) of a query vector X_{query} (*1*-by-*d*) is predicted as:

$$Y_{predict} = X_{query} B_{PLS} + \varepsilon \quad (12)$$

where ε is the intercept. For binary classification, a threshold of 0.5 is commonly applied for category decision. For multiclass problem, the category decision follows the maximum assignment criterion [22, 23] where the column of maximum value in vector $y_{predict}$ returned is the predicted category.

2.3 Local PLS-DA

Local PLS-DA employs adjacent references of a query instead of global data during PLS modelling thus can be viewed as a Just-in-Time (JIT) approach. Typical distance functions such as Euclidean, Manhattan, cosine and correlation distance have been widely used to select nearest references. Despite the fact that distance function plays a key role in the neighborhood selection, the effect of different distance functions towards datasets as well as learning tasks are varies. For example, Euclidean distance performs less well than cosine and correlation distance on the alignment of mass spectrometry data [24, 25]. Also, Manhattan or fractional distance can provide better discriminations when dimensionality increases [19, 20]. Such findings motivate the embedment of different distance functions in local PLS-DA. Given data vectors $X_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ and $X_j = (x_{j1}, x_{j2}, \cdots, x_{jd})$ in

data matrix X_{nd} , following distance functions are applied in our work:

- Euclidean distance:

$$D_{ij}^2 = (X_i - X_j)(X_i - X_j)' \quad (13)$$

- Manhattan distance:

$$D_{ij} = \sum_{k=1}^{d} | x_{ik} - x_{jk} | \quad (14)$$

- Cosine distance:

$$D_{ij} = 1 - \frac{X_i X_j^{'}}{\sqrt{(X_i X_i^{'})(X_j X_j^{'})}} \quad (15)$$

- Correlation distance:

$$D_{ij} = 1 - \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{\mathrm{var}(X_i)\,\mathrm{var}(X_j)}} \quad (16)$$

- Fractional distance:

$$D_{ij} = (\sum_{k=1}^{d} | x_{ik} - x_{jk} |^p)^{1/p},\ p \in (0,1) \quad (17)$$

In this paper, p value in fractional distance is set to 1/2, 1/3 and 1/5. For a query X_{query} , Nearest Neighbors (NNs) is firstly selected as local references X_{local} with corresponding categories of dummy matrix Y_{local} . If all of the local references belong to a same category, the query is directly attributed to that category. Otherwise given a specified LVs number, the PLS regression coefficient and intercept can be obtained by SIMPLS algorithm. With (12), the category of X_{query} will be predicted. The optimal parameters of LVs and NNs are usually identified by cross validation. For LVs, it will not over the dimensionality of the data. Usually the performance is not affected significantly if the number of LVs is greater than 5 [26]. Reference [27] set the number of LVs considers the eigen-decomposition of the between-class scatter matrix in PLS-DA. The range of NNs can be empirically set which takes into account the specificities of data. For example, fewer NNs may cause under fitting while too many NNs will resemble local PLS-DA to global PLS-DA.

3. Experiments

To gauge the performance of local PLS-DA coupled with various distance functions, 10 datasets of multi-class and high-dimensionality have been selected. The analysis of the obtained results and classification accuracy take account of parameter setting; that is the number of NNs used to build local neighborhoods and number of LVs in PLS models.

3.1 Datasets

- 8 UCI repository datasets: from the UCI repository [28], 8 datasets have been selected which cover data types of high-dimensionality, multi-class and imbalance. The basic information about these datasets is shown in **Table 1**.

Table 1 Information on 8 datasets from UCI repository and the ranges of parameters in the first experiment

Datasets	Instances	Attributes	Categories	LVs	NNs
Breast Tissue	106	10	2	1-9	10-50
Ecoli	336	8	8	1-7	20-60
Glass	214	9	6	1-9	20-60
Ionosphere	351	34	2	1-7	30-80
Parkinson's	197	23	2	1-10	30-70
PLRX	182	13	2	1-9	30-70
Sonar	208	60	2	1-9	10-40
Wine	178	12	3	1-12	20-60

- Near-Infra Red (NIR) cookie dataset: This dataset was initially generated as part of an experiment to test the feasibility of NIR spectroscopy to measure the composition of 4 responses: namely fat, sucrose, dry flour and water [29, 30]. It comprises 700 variables captured at wavelengths ranging from 1100 to 2498 nm with a distance interval of 2 nm. There are 40 samples in the training set and 32 samples in the test set. We simplify the regression task to classification by setting cut-off values which divide selected responses of fat and water into binary groups; cut-off values of 18.31% for the mean fat response and 14.60% for the mean water response are used to discretize fat and water responses into binary categories as shown in **Fig.1**.

- ARCENE dataset: This dataset was initially used in the Neural Information Processing Systems (NIPS) 2003 feature selection challenge [31]. It is obtained by mass-spectrometer and aims to identify cancer from normal patterns. There are 10000 randomly sorted features in the ARCENE dataset among which 3000 features with no predictions are added as 'distractors'. As the labels of test samples are not available, we have selected the validation set as test data in Task 1 and then have inverted this procedure in Task 2. The ratio of training to test samples in both procedures is 1:1 as shown in **Table 2**.

Table 2 Sample distribution of ARCENE dataset

ARCENE	Positive	Negative	Total
Training set	44	56	100
Validation set	44	56	100
Test set (N/A)	310	390	700
All	398	502	900

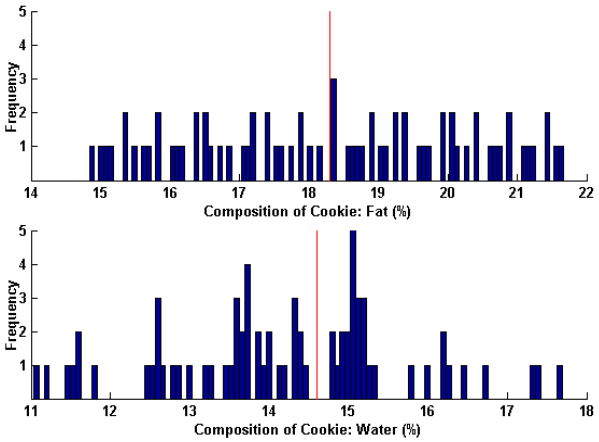

Fig.1. Histogram of fat and water response (%) in NIR-Cookie dataset. Vertical lines in red discretize the dataset into binary categories with cut-off values of 18.31% and 14.60% in fat and water response, respectively.

3.2 Experimental settings

In the experiments on 8 UCI datasets, a 10 fold cross-validation approach was applied to obtain an average of validation accuracy of PLS based classifications on 8 UCI datasets. We set the range of parameters LVs and NNs depending on the specified dataset. Basically for classification purposes, both parameters are determined by cross-validation which returns a LVs-by-NNs validation accuracy matrix. Previous work suggested that the optimal number of LVs in PLS-DA is usually less than twenty [32]. Another parameter NNs is set by using a distance function which adjusts the local PLS-DA model towards a given query point. Intuitively, too many NNs will reduce the local property while too few NNs barely provide enough distinctions. Also, the NNs can be empirically fixed to several values and ranges in order to test the performance of local PLS [10, 12]. In our experiment, two ranges of parameters (LVs and NNs) are selected and listed in **Table 1**.

The second and third experiment provides a case study of using local PLS-DA to classify small sample sized chemical spectroscopy data. As training and testing sets have been assigned, a 10-fold cross validation step is firstly operated within training set to identify the optimal LVs and NNs. Then the returned parameters are used for modelling and classification. The range of LVs in local PLS-DA is from 1 to 8 for both experiments, while the range of NNs is 10:35 and 10:50 with a step of 1, respectively in NIR cookie and ARCENR datasets. The mean value of NNs within a range is estimated by the number of points in potential groups after latent projections. For example in ARCENE dataset, 100 training points can be divided into three groups in 2 or 3 dimensional latent spaces (**Fig.2**). The searching range of NNs is set to avoid too many and too few modelling references. Additionally, these experiments aim to show that, given reasonable ranges of both parameters in local PLS-DA, the performance has already exceeded global PLS-DA and the outcomes vary when coupled with different distance functions.

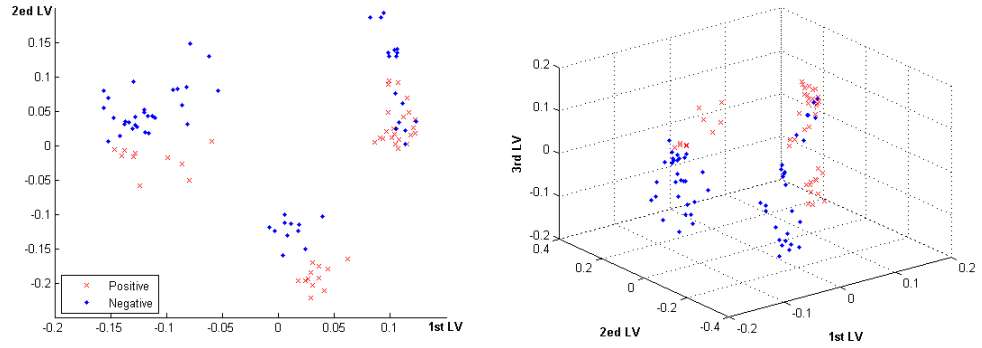

Fig.2. 2D and 3D scatter plots of PLS latent variables in representing ARCENE training samples in Task-1

4. Results and discussion

The overall results of 8 UCI datasets using PLS-DA and local PLS-DA are given in **Table 3**. In most of the datasets, local PLS-DA drastically exceeds global PLS-DA regardless of the distance function used. Usually, when LVs in local PLS-DA is smaller than the optimal number of LVs in global PLS-DA, the performance has already exceeded the global ones. Given a query point in **Fig.6** (197th sample in Sonar dataset), global and local PLS-DA project neighboring and overall references, respectively into the space spanned by two latent variables. Global PLS-DA achieves an accuracy of 82.19% using 7 LVs, while local PLS-DA can be 7% higher using 3 LVs only. The boundary between two categories in global PLS-DA is less distinctive compared to local PLS-DA.

Within local PLS-DA, fractional and cosine distance dominate the highest accuracies achieved. Euclidean distance always performs less well than other distance functions. For example in **Fig.4** and **5**, a high-level of accuracies is obtained by the proposed fractional distance in Breast Tissue and ECOLI datasets given the searching range of LVs and NNs. Similarly, it can be seen in **Fig.6** that cosine distance globally outperforms Euclidean distance in ionosphere dataset. Further, the cosine and correlation distance have attained comparable validation accuracies and parameters value, for instance in the Parkinson's (**Fig.4**) and PLRX (**Fig.5**) datasets.

Table 3 Average accuracies of 10-fold cross validation obtained by Global and local PLS-DA with optimal parameters. Local PLS-DA are based on Euclidean, Manhattan, Cosine, Correlation and Fractional (1/2, 1/3 and 1/5) distance.

PLS-DA		Breast Tissue			ECOLI			Glass		
		Accuracy	LVs	NNs	Accuracy	LVs	NNs	Accuracy	LVs	NNs
Global		0.8300	3	N/A	0.8542	5	N/A	0.6203	8	N/A
Local	Euclidean	0.8873	7	16	0.8782	2	41	0.7290	5	22
	Manhattan	0.8955	4	13	0.8782	1	34	0.7478	5	22
	Cosine	0.8873	1	11	0.8722	1	33	**0.7483**	8	27
	Correlation	0.8782	1	11	0.8751	3	53	0.7385	8	27
	1/2	**0.9145**	6	15	0.8812	5	56	0.7245	4	23
	1/3	0.8955	3	17	0.8842	5	57	0.7240	4	42
	1/5	0.8964	3	15	**0.8902**	6	43	0.7338	4	27
PLS-DA		Ionosphere			Parkinson's			PLRX		

PLS-DA		Sonar			Wine					
		Accuracy	LVs	NNs	Accuracy	LVs	NNs	Accuracy	LVs	NNs
Global		0.8718	6	N/A	0.8661	7	N/A	0.7143	1	N/A
Local	Euclidean	0.9487	2	79	0.9071	8	37	0.7143	1	63
	Manhattan	0.9460	5	77	0.9221	7	47	0.7091	1	46
	Cosine	0.**9601**	3	78	0.9076	10	54	0.**7254**	1	45
	Correlation	0.9515	3	79	0.9076	10	54	0.**7254**	1	47
	1/2	0.9315	1	32	0.9174	8	40	0.7088	1	57
	1/3	0.9346	2	54	0.**9224**	9	44	0.7085	2	54
	1/5	0.9317	1	46	0.9171	8	30	0.7143	2	59
PLS-DA		Sonar			Wine					
		Accuracy	LVs	NNs	Accuracy	LVs	NNs			
Global		0.8219	7	N/A	0.9775	8	N/A			
Local	Euclidean	0.8938	3	22	0.9889	11	59			
	Manhattan	0.8940	7	11	0.**9944**	10	59			
	Cosine	0.8798	3	34	0.**9944**	12	44			
	Correlation	0.8938	3	22	0.9886	11	59			
	1/2	0.8938	8	18	0.9778	9	23			
	1/3	0.8990	3	13	0.9778	11	27			
	1/5	0.**9038**	3	10	0.9889	12	59			

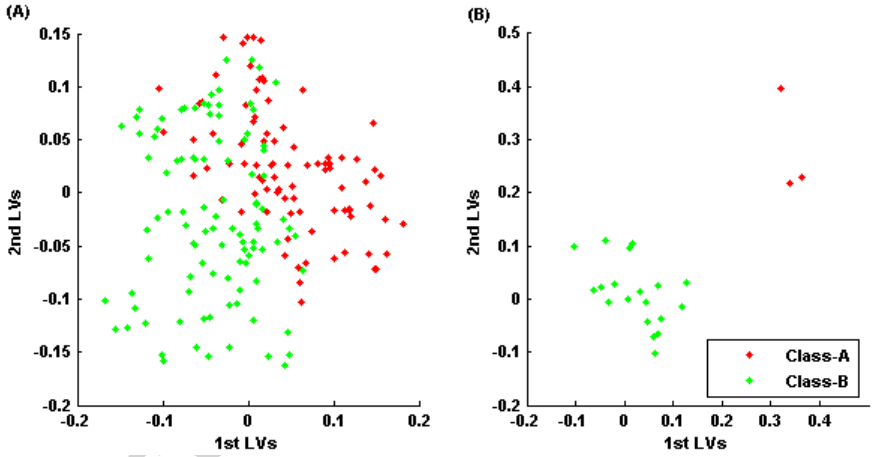

Fig.3. Latent projections of training data on the classification of 197th sample in Sonar dataset: global (A) and local Euclidean (B) PLS-DA. The LVs in (A) and (B) is 7 and 3, respectively. The NNs in (B) is 22.

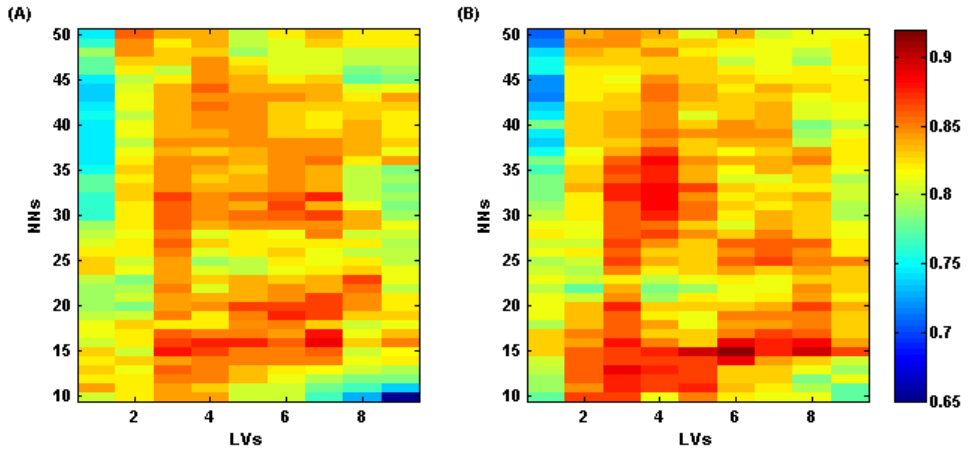

Fig.4. Average accuracies of local PLS-DA with LVs and NNs represented in color map: Euclidean (A) and fractional 1/2 (B) distance in Breast Tissue dataset.

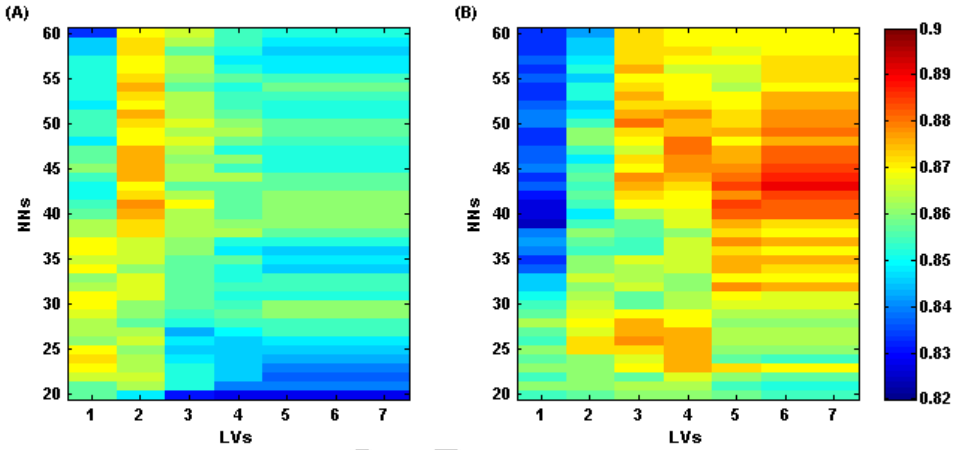

Fig.5. Average accuracies of local PLS-DA with LVs and NNs represented in color map: Euclidean (A) and fractional 1/5 (B) distance in ECOLI dataset.

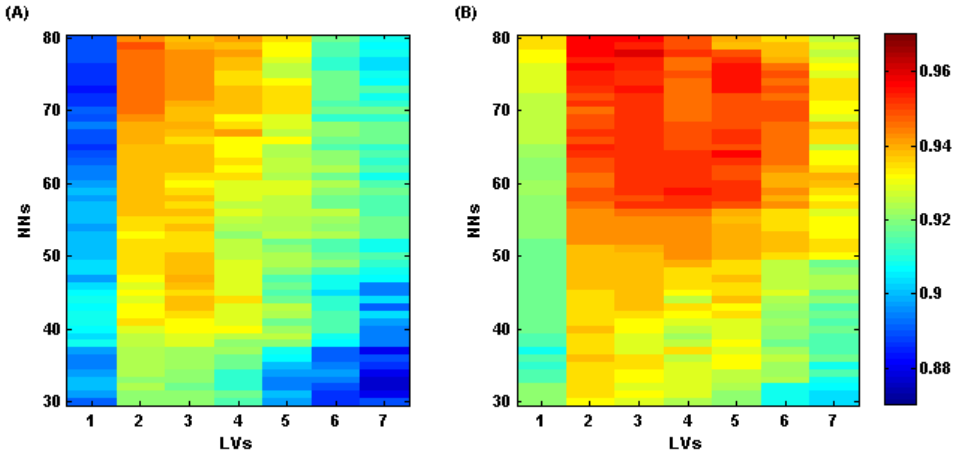

Fig.6. Average accuracies of local PLS-DA with LVs and NNs represented in color map: Euclidean (A) and cosine (B) distance in ionosphere dataset.

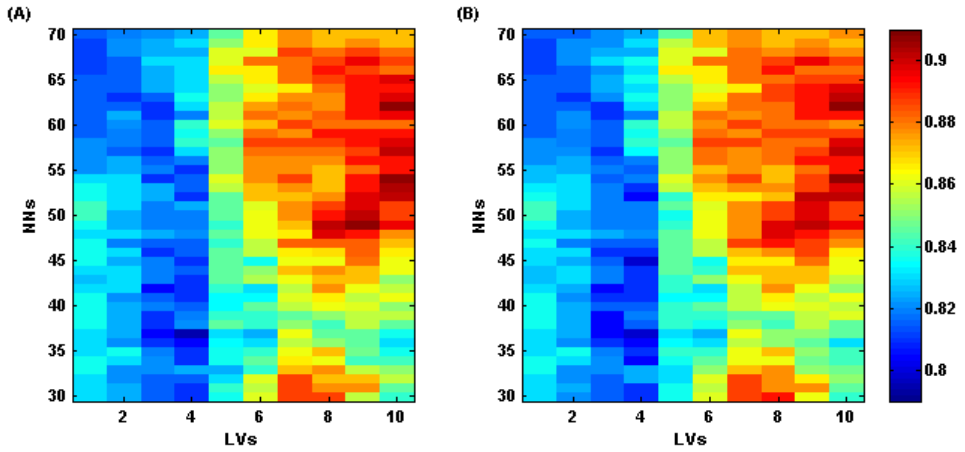

Fig.7. Average accuracies of local PLS-DA with LVs and NNs represented in color map: cosine (A) and correlation (B) distance in Parkinson's dataset.

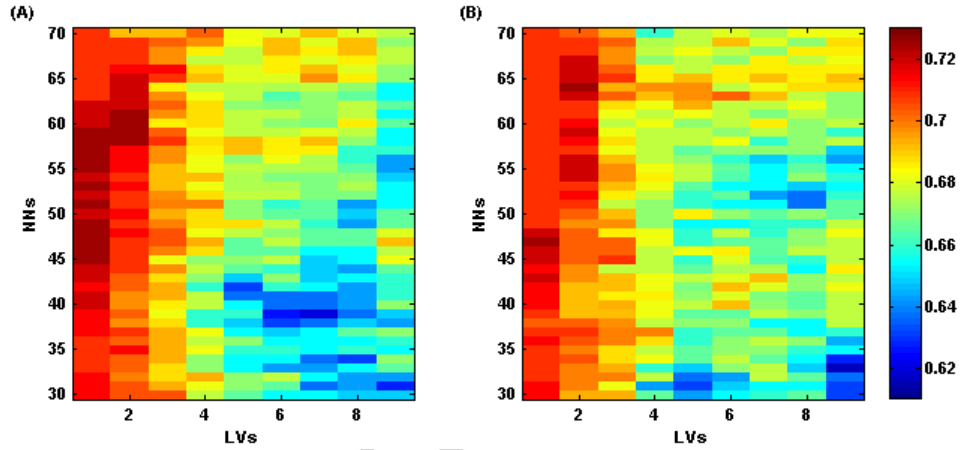

Fig.8. Average accuracies of local PLS-DA with LVs and NNs represented in color map: cosine (A) and correlation (B) distance in PLRX dataset.

The second experiment deals with the classification of NIR-Cookie dataset. The average results of 10-fold cross validation within training set are presented as a 3D plot in **Fig.9** where a peak area indicates high classification accuracy. For instance, when the number of NNs is around 30 and the number of LVs is 5, local PLS-DA based on Euclidean distance reaches near 98% accuracy for the validation of fat response as depicted in **Fig.9-A**. The same observation and result are valid for the case where the Manhattan distance is used as shown in **Fig.9-B**. The validation and classification results of both parameters in local PLS-DA are given in the boxplot of **Fig.10**. There are 208 pairs of accuracy (LVs×NNs) in each distance function; the edges of the box are the 25th and 75th percentiles while the central mark is the median value. The whiskers extend to the most extreme data points not considered to be outliers; the outliers are plotted individually. The horizontal red dot line in **Fig.10** shows the highest accuracy that global PLS-DA can achieve. It can be seen that local PLS-DA based on the first five distances in Fig.4 reach the highest accuracy of 98%; among these distances the Euclidean distance achieves slightly lower performance than the other four at the median central mark, 25th and 75th percentiles edges. Further, the classification results using selected NNs and LVs parameters in validation phase are shown in **Table 4**. A typical case is the validation and classification of water response based on Euclidean distance. As shown in **Fig.9-C**, the highest validation accuracy attains a

local peak value of 89.67% where NNs and LVs are 20 and 6 respectively. However, in classification phase 78.13% accuracy can be obtained by using such parameters. From the second experiment, when dimensionality becomes higher in NIR cookie dataset, the highest accuracy obtained by global PLS-DA is either comparable or above the 25th percentile of local PLS-DA; however it tends to not attain the highest accuracy, in particular in local PLS-DA based on fractional distance.

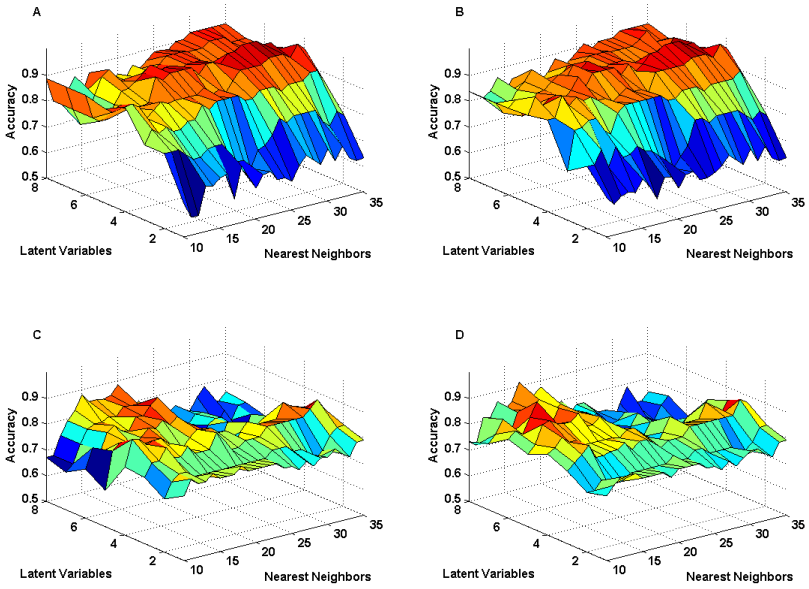

Fig.9. Average accuracies (10-fold cross validation) of local PLS-DA with Latent Variables and Nearest Neighbors represented in 3D plots: Euclidean (**A**) and Manhattan (**B**) distance in fat response & Euclidean (**C**) and fractional (p = 1/3) distance (**D**) in water response.

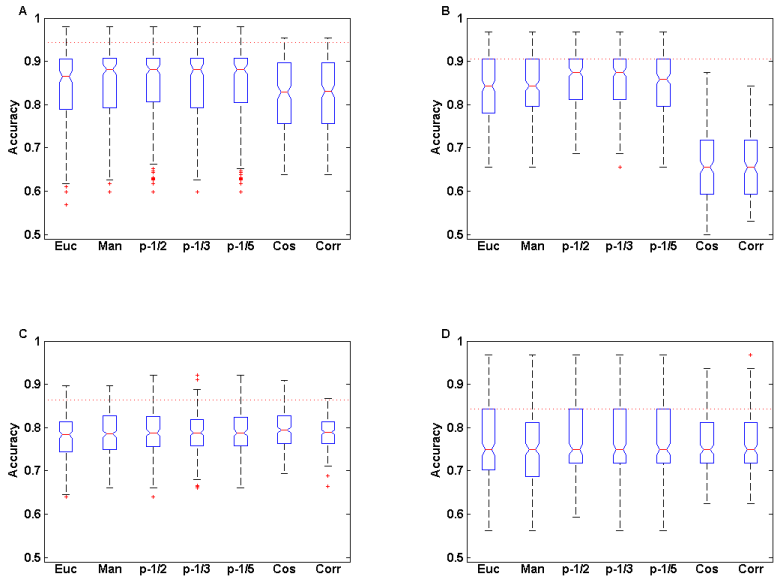

Fig.10. Comparisons within local PLS-DA based on Euclidean (Euc), Manhattan (Man), fractional (p = 1/2, 1/3 and 1/5), Cosine

(Cos) and Correlation (Corr) distance in NIR cookie dataset: validation (**A**) and classification (**B**) in fat response & validation (**C**) and classification (**D**) in water response. The red dot line (horizontal) represents the highest accuracy obtained by global PLS-DA.

Table 4 Comparison of global PLS and local PLS-DA performance on NIR cookie dataset: classification accuracy (%) of Fat and Water categories.

Classification (%)		**PLS-DA**	**Local PLS-DA**						
			Euclidean	Manhattan	Fractional			Cosine	Correlation
					p=1/2	*p*=1/3	*p*=1/5		
Fat	Accuracy	90.63	93.75	93.75	96.88	96.88	96.88	84.38	84.38
Parameters	LVs	4&5	5	5	5	5	5	4	4
	NNs	N/A	30	30	30	30	30	35	35
Water	Accuracy	84.38	93.75	93.75	96.88	96.88	93.75	86.33	86.33
Parameters	LVs	14	4	4	4	4	4	3	3
	NNs	N/A	32	32	32	32	32	30	30

The third experiment has been carried out on the ARCENE dataset. The accuracies of global PLS-DA with LVs ranging from 1 to 20 are shown in **Fig.11**. In task-1, the accuracy starts from a minimum value of 0.60 and peaks at 0.86 when the number of LVs equals 3. With fluctuations before 12 LVs, it finally stabilizes at 0.82. In task-2, the PLS-DA accuracy begins at 0.63 and decreases to a minimum value of 0.55 in next LVs. Then it increases to a maximum of 0.90 when LVs equals to 8. After that, the accuracy appears to show the same trend as with task-1 but is 8% higher. There are 328 pairs of accuracy (LVs × NNs) presented in each of the distance function during validation. We calculate the mean accuracies of LVs with variation of NNs and select corresponding NNs returned from top 3 out of 41 mean accuracies to define the parameters in classification phase. By using such parameters, the relatively optimal classification results are achieved as shown in **Table 5**. The Euclidean distance generally has lower performance than other distances in the validation of local PLS-DA; however it maintains a same level compared to other distances in classification phase. Most of the local PLS-DA results exceed global PLS-DA with fewer or equal LVs. In particular, in task-1 if LVs and NNs are equals to 1 and 27 for local PLS-DA based on Manhattan distance, respectively, an accuracy of 0.88 can be obtained which is already above 0.86 of global PLS-DA. Similar case also exists in Euclidean and fractional for p = 1/3 distance. Another optimal parameter NNs is ranging from 22 to 32 in task-1 which has connections with the size of clusters shown in **Fig.2**.

Further, overall PLS-DA validation and classification result of both tasks are calculated and presented in **Fig.1**2. The outliers shown in red dot under the whisker are identified based on the analysis of boxplot [33]. The number of identified outliers is less than 20 in each of the distance function. The overall classification accuracies are visualized in 3D in **Fig.13**; such outliers correspond to a lower region, for instance, when LVs and NNs are around 1 and 50, respectively. As previously observed in the second experiment, the highest accuracy obtained using PLS-DA is usually above the 25th percentile of local PLS-DA however it is still lower than the highest accuracy of local PLS-DA in both validation and classification phases. Among different fractional distance based local PLS-DA, Manhattan and fractional distance can obtain better results than Euclidean distance in the evaluation of

whisker and box. A typical case is the validation in task-1, Manhattan and fractional (p = 1/2 and 1/3) can exceed Euclidean distance in maximum, medium, 25th and 75th percentiles. Another two distance functions, cosine and correlation yield a similar result in the classification of both tasks (**Fig.12, B&C**). In particular, they attain a peak value of 93% in task-2. By comparing the parameters in classification phase shown in **Table 6**, we can see that the optimal LVs of local PLS-DA in both tasks are usually smaller than that of global PLS-DA.

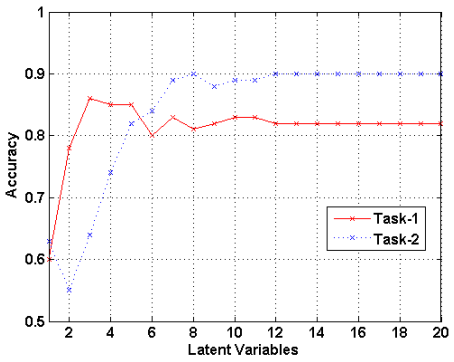

Fig.11. Classification accuracy of global PLS-DA with the selection of latent variables, **Task-1:** training set is original training data while test set is original validation data; **Task-2:** training set is original validation data while test set is original training data.

Table 5 The average validation and classification results (%) of local PLS based on 7 different distances: Firstly, the mean accuracies of LVs with variation of NNs are calculated. Then corresponding NNs returned from top 3 out of 41 mean accuracies are used to define the parameters in classification.

Local PLS-DA (%)		Euclidean	Manhattan	Fractional			Cosine	Correlation
				p=1/2	p=1/3	p=1/5		
Task - 1	Validation	86.36	86.43	86.67	87.21	86.90	85.71	86.81
	Classification	89.00	89.00	87.00	86.00	88.00	89.00	89.00
Parameters	LVs	4	7	3	5	2	4	6
	NNs	25	27	24	32	24	22	26
Task - 2	Validation	81.88	81.92	81.66	82.54	82.55	81.96	81.39
	Classification	91.00	91.00	90.00	89.00	90.00	91.00	90.00
Parameters	LVs	5	4	3	3	3	4	5
	NNs	33	29	27	23	23	35	32

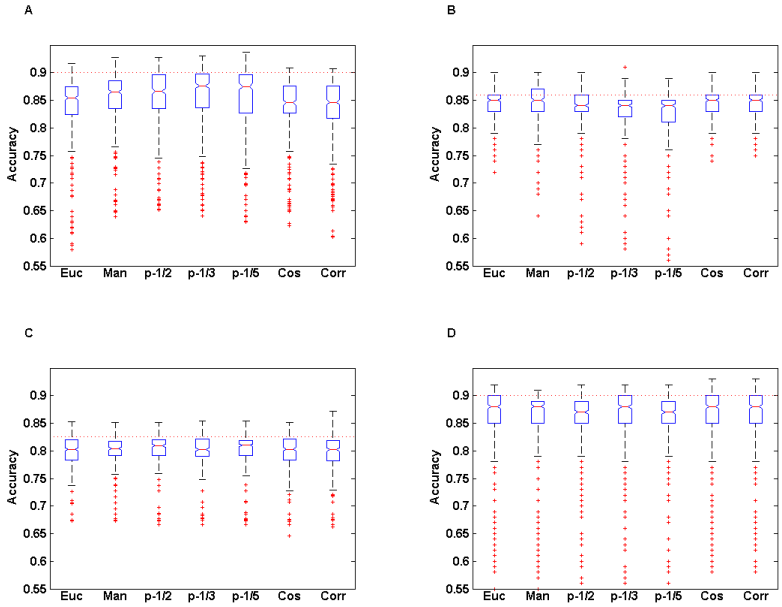

Fig.12. Comparisons within local PLS-DA based on Euclidean (Euc), Manhattan (Man), fractional (p = 1/2, 1/3 and 1/5), Cosine (Cos) and Correlation (Corr) distance in ARCENE dataset: validation (**A**) and classification (**B**) in task-1 & validation (**C**) and classification (**D**) in task-2. The red dot line (horizontal) represents the highest accuracy obtained by global PLS-DA.

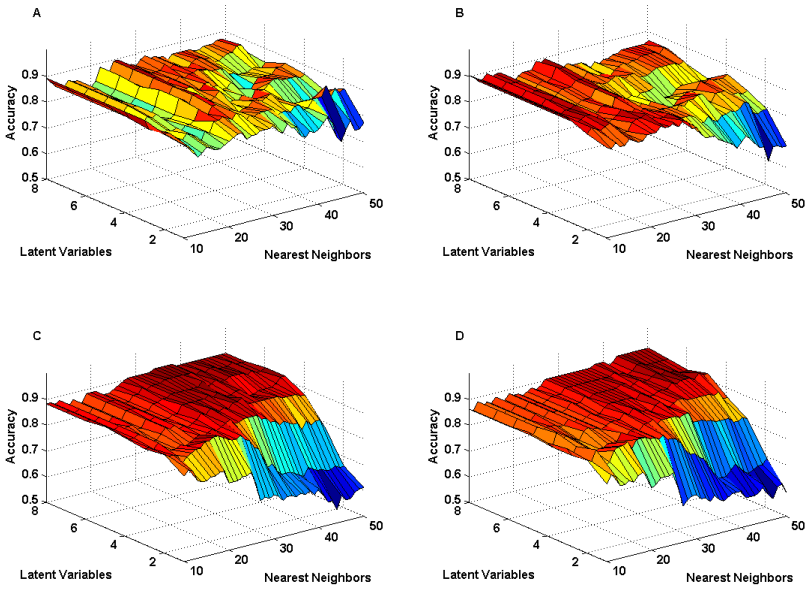

Fig.13. Classification accuracy of local PLS-DA with Latent Variables and Nearest Neighbors represented in 3D plots: Euclidean (**A**) and Manhattan (**B**) distance in Task-1 & Euclidean (**C**) and Cosine distance (**D**) in Task-2.

Table 6 Highest accuracies (%) obtained by different local PLS-DA with minimum number of LVs and NNs in the ARCENE dataset.

Local PLS-DA (%)		Euclidean	Manhattan	Fractional			Cosine	Correlation
				p=1/2	p=1/3	p=1/5		
Task-1	Accuracy	90.00	90.00	90.00	91.00	89.00	90.00	90.00
Parameters	LV	2	3	3	2	2	4	2
	NN	46	10	11	10	16	23	23
Task-2	Accuracy	92.00	91.00	92.00	92.00	92.00	93.00	93.00
Parameters	LV	4	4	5	5	5	5	5
	NN	36	29	45	35	40	48	47

Considering the experimental results, the following results and issues have been identified:

- The local PLS-DA can significantly outperform global PLS-DA in most of the 8 UCI datasets. Compared with the global view of PLS-DA, the local PLS-DA only projects a group of similar data points to a low-dimensional latent space in which the intrinsic latent structure of local data is preserved. Also, the LVs in local PLS-DA is usually smaller than that in global ones.
- As a JIT approach, local PLS-DA can analyse query data with effectiveness in terms of accuracy and flexibility of neighbourhood adapting; however it is highly reliant upon distance functions. In our experiment, the combination of fractional distance and local PLS-DA usually outperforms the Euclidean ones in 8 UCI datasets. Cosine distance is also preferable in datasets such as Ionosphere and PLRX. Therefore, distance functions need adoptions towards different data type.
- Noteworthy, the optimal parameters selected in the validation step may not yield the highest classification accuracy. This reflects a real-world challenge in the analysis of some low resolution spectroscopy data when training and test samples are collected under two different conditions. As such, the classification result between two datasets may be drastically lower than the validation results within the same dataset.
- Local PLS-DA cannot yield a global model. Moreover, a potential issue of local PLS-DA is additional parameter NNs is involved, thus distance between data need to be calculated and sorted. Also, individual model is established for each query data. These will lead to high modelling time when dimensionality and sample-size are extremely large.

5. Conclusion

Local PLS-DA can construct classification models in local sample space. They not only inherit the advantage of PLS in handling small-sample-size problems, they also avoid the influence from noise samples if nearest neighbors are properly selected. In this paper, we combine local PLS-DA and non-Euclidean distance in order to gain an insight, in particular, to find out how different distance functions perform on low- and high-dimensional data.

Extensive experiments have been conducted using various distance functions with a local PLS-DA. The obtained results show that local PLS-DA based on fractional and cosine distance is preferred to the

Euclidean distance in the classification of high-dimensional data. Furthermore, better classification performance usually can be achieved when the number of LVs in local PLS-DA is smaller than that in the global PLS-DA. This shows that the intrinsic dimension in local latent space is simplified yet effective in improving the classification accuracy compared with global one.

Future work will optimize the number of NNs by effective noise detection techniques with better representation of local structures and computational speed.

References

[1] Saul, L. K. L., & Roweis, S. S. T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, *4(1999)*, 119–155. http://doi.org/10.1162/153244304322972667

[2] Ingram, S., & Munzner, T. (2015). Dimensionality reduction for documents with nearest neighbor queries. *Neurocomputing*, *150(PB)*, 557–569. http://doi.org/10.1016/j.neucom.2014.07.073

[3] Zhuo, L., Cheng, B., & Zhang, J. (2014). A comparative study of dimensionality reduction methods for large-scale image retrieval. *Neurocomputing 141*, 202–210. http://doi.org/10.1016/j.neucom.2014.03.014

[4] Wold, H. (1985). Partial Least Squares. *International Journal of Cardiology*, *147*(2), 581–591. http://doi.org/10.1016/j.ijcard.2010.12.060

[5] Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173. http://doi.org/10.1002/cem.785

[6] Sugiyama, M. (2007). Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. *Journal of Machine Learning Research*, *8*, 1027–1061. Retrieved from http://portal.acm.org/citation.cfm?id=1248659.1248694

[7] Pedagadi, S., Orwell, J., Velastin, S., & Boghossian, B. (2013). Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3318–3325). http://doi.org/10.1109/CVPR.2013.426

[8] Verleysen, M., & François, D. (2005). The Curse of Dimensionality in Data Mining. In *Computational Intelligence and Bioinspired Systems* (pp. 758–770). http://doi.org/10.1007/11494669_93

[9] Kim, S., Kano, M., Nakagawa, H., & Hasebe, S. (2011). Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International Journal of Pharmaceutics*, *421*(2), 269–274. http://doi.org/10.1016/j.ijpharm.2011.10.007

[10] Bevilacqua, M., & Marini, F. (2014). Local classification: Locally weighted-partial least squares-discriminant analysis (LW-PLS-DA). *Analytica Chimica Acta*, *838*, 20–30. http://doi.org/10.1016/j.aca.2014.05.057

[11] Bottou, L., & Vapnik, V. (1992). Local Learning Algorithms. *Neural Computation*, *4*(6), 888–900. http://doi.org/10.1162/neco.1992.4.6.888

[12] Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., & Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry*, *68*, 337–347. http://doi.org/10.1016/j.soilbio.2013.10.022

[13] Tian, X., Wu, Z., & Chng, E. S. (2013). Local partial least square regression for spectral mapping in voice conversion. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013*. http://doi.org/10.1109/APSIPA.2013.6694332

[14] Hazama, K., & Kano, M. (2015). Covariance-based locally weighted partial least squares for high-performance adaptive modeling. *Chemometrics and Intelligent Laboratory Systems*, *146*, 55–62. http://doi.org/10.1016/j.chemolab.2015.05.007

[15] He, X., & Niyogi, P. (2004). Locality preserving projections. *Neural Information Processing Systems*, *16*, 153. http://doi.org/10.1.1.19.9400

[16] He, K., Cheng, H., Du, W., & Qian, F. (2014). Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. *Chemometrics and Intelligent Laboratory Systems*, *134*, 79–88. http://doi.org/10.1016/j.chemolab.2014.03.007

[17] Shigemori, H., Kano, M., & Hasebe, S. (2011). Optimum quality design system for steel products through locally weighted regression model. In *Journal of Process Control* (Vol. 21, pp. 293–301). http://doi.org/10.1016/j.jprocont.2010.06.022

[18] Nakagawa, H., Tajima, T., Kano, M., Kim, S., Hasebe, S., Suzuki, T., & Nakagami, H. (2012). Evaluation of infrared-reflection absorption spectroscopy measurement and locally weighted partial least-squares for rapid analysis of residual drug substances in cleaning processes. *Analytical Chemistry*, *84*(8), 3820–3826. http://doi.org/10.1021/ac202443a

[19] Aggarwal, C. C., Hinneburg, A., & Keim, D. a. (2001). On the surprising behavior of distance metrics in high dimensional space. *Database Theory – ICDT 2001*, 420–434. doi:10.1007/3-540-44503-X_27

[20] Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? *Database Theory—ICDT'99*, 217–235. http://doi.org/10.1007/3-540-49257-7_15

[21] de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263. http://doi.org/10.1016/0169-7439(93)85002-X

[22] Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790–3798. http://doi.org/10.1039/C3AY40582F

[23] Néstor F. Pérez, Joan Ferré, Ricard Boqué. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. Chemometrics and Intelligent Laboratory Systems, 95(2), 122–128. http://doi.org/10.1016/j.chemolab.2008.09.005

[24] Kim, S., Fang, A., Wang, B., Jeong, J., & Zhang, X. (2011). An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure. *Bioinformatics*, *27*(12), 1660–1666. http://doi.org/10.1093/bioinformatics/btr188

[25] Kim, S., & Zhang, X. (2013). Comparative analysis of mass spectral similarity measures on peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry. *Computational and Mathematical Methods in Medicine*, *2013*. http://doi.org/10.1155/2013/509761

[26] Nguyen, D. V, & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, *18*(1), 39–50. http://doi.org/10.1093/bioinformatics/18.1.39

[27] Qu, H.-N., Li, G.-Z., & Xu, W.-S. (2010). An asymmetric classifier based on partial least squares. *Pattern Recognition*, *43*(10), 3448–3457. http://doi.org/10.1016/j.patcog.2010.05.002

[28] Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository. *University of California Irvine School of Information*. Retrieved from http://www.ics.uci.edu/~mlearn/MLRepository.html

[29] Osborne, B. G., Fearn, T., Miller, A. R., & Douglas, S. (1984). Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, *35*(1), 99–105. http://doi.org/10.1002/jsfa.2740350116

[30] Brown, P., Fearn, T., & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, *96*(July 2015), 398–408. http://doi.org/10.1198/016214501753168118

[31] Guyon, I., Gunn, S., Ben-Hur, A., & Dror, G. (2005). Result Analysis of the NIPS 2003 Feature Selection Challenge. In Advances in Neural Information Processing Systems 17 (pp. 545–552).

[32] Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790–3798. http://doi.org/10.1039/C3AY40582F

[33] McGill, R., Tukey, J. W., & Larsen, W. a. (1978). Variations of Box Plots. *The American Statistician*, *32*(1), 12–16. http://doi.org/10.2307/2683468

Weiran Song received his BSc degree in Applied Mathematics in 2010 from Hebei University, China and MSc degree in Mathematics from Beijing University of Technology in 2013, China. He is currently working towards the Ph.D. degree in Computer Science at Ulster University. His research interests are data mining and pattern recognition.

Hui Wang is Professor of Computer Science at Ulster University. His research interests are machine learning, logics and reasoning, combinatorial data analytics, and their applications in image, video, spectra and text analysis. He has over 200 publications in these areas.

Paul Maguire is Professor of Plasmas and Nanofabrication at Ulster University. His main research interests lie in the development and integration of microplasma, microfluidic and microfabrication techniques for application in new clinical and environmental sensors devices, biocompatible nanomaterials coatings, 3rd generation photovoltaic solar cell devices, biological and food treatments and nanomaterials for industry.

Omar Nibouche received his BEng degree in Electronic Engineering form the Polytechnic School of Algiers and PhD degree in Computer Science from Queen's University Belfast. He is a lecturer in computing in the School of Computing and Mathematics, Ulster University at Jordanstown. His research interests include machine learning, applications of artificial intelligence, computer vision and biometrics.