EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

A Comparative Analysis of Feature Extraction Methods for Classifying Colon Cancer Microarray Data

M.O Arowolo[1,]*, R.M. Isiaka[1], S.O. Abdulsalam[1], Y.K. Saheed[2] and K.A. Gbolagade[1]

[1,] Department of Computer Science, College of Information and Communication technology,
Kwara State University, Malete, Nigeria.
[2,] Department of Physical Sciences, Al-Hikmah University, Ilorin, Kwara State, Nigeria.

Abstract

Feature extraction is a proficient method for reducing dimensions in the analysis and prediction of cancer classification. Microarray procedure has shown great importance in fetching informative genes that needs enhancement in diagnosis. Microarray data is a challenging task due to high dimensional-low sample dataset with a lot of noisy or irrelevant genes and missing data. In this paper, a comparative study to demonstrate the effectiveness of feature extraction as a dimensionality reduction process is proposed, and concludes by investigating the most efficient approach that can be used to enhance classification of microarray. Principal Component Analysis (PCA) as an unsupervised technique and Partial Least Square (PLS) as a supervised technique are considered, Support Vector Machine (SVM) classifier were applied on the dataset. The overall result shows that PLS algorithm provides an improved performance of about 95.2% accuracy compared to PCA algorithms.

Keywords: Dimensionality Reduction, Principal Component Analysis, Partial Least Square, Support Vector Machine.

Received on 12 July 2017, accepted on 09 September 2017, published on 25 September 2017

doi: 10.4108/eai.25-9-2017.153147

1. Introduction

Dimensionality reduction is a very helpful, important and necessary tool in the expression of microarray datasets. It endeavours to trim down, recognize and illustrates the collection of unified datasets by transforming a high-dimensional dataset into a lower dimensional dataset which signifies the most significant variables that triggers the distinctive data. This significant and essential tool attracts numerous researchers working in the aspect of bioinformatics and deals with gene expression datasets to work on the dimensionality reduction [1], [2]. Several methods for dimension reduction exist, but none is confirmed the best method for all circumstances due to the reason that at the time of the processing some information is lost [3]. To improve the performance of dataset, feature extraction as a universal method of dimension reduction is considered. In Feature extraction method, the original high-dimensional feature space is estimated on to low-dimensional feature space, for typical microarray data analysis the training sample size is always limited. Due to classification algorithms may be short of efficiency or even fail in high dimensional microarray data analysis, dimension reduction is a good choice to variable selection in order to overcome the dimensionality problem; it uses a little quantity of features to substitute a feature subset containing well-built correlations in the original data [4]. Quite a lot of feature extraction algorithms and techniques have been proposed in literature, one of the most popular and widely used techniques is PCA [1]. PCA is an unsupervised method and an effective tool but it is not efficient for high dimensional and complex dataset, due to the fact that it cannot retrieve precisely the true latent variables of complex and supervised datasets [5], data in a very high dimensional space often exists in a lower dimension. With this kind of data, the intrinsic supervised structure could

[*] Corresponding author. Email:olliray2002@yahoo.com

not be found through an unsupervised feature extraction technique. Another drawback of PCA is that the size of the covariance matrix is proportional to the dimensionality of the data-points. In order to overcome the drawback of unsupervised feature extraction in a very high dimensional dataset, several supervised feature extraction methods have been developed. An improvement of supervised algorithm is Local Linear Embedding (LLE) [1], it is efficient and powerful for dimensionality reduction among the other algorithms [5], [6], [7]. Local Tangent Space Analysis (LTSA) is another nonlinear dimensionality reduction technique that describes local properties of the high-dimensional data using the local tangent space of each data point [8]. These techniques have been successfully applied on microarray data.

In this paper, a supervised feature extraction algorithm for dimensionality reduction is proposed to handle the curse of dimensionality of microarray data. PLS is a proposed algorithm for supervised feature extraction. The experiments show PLS outperforms PCA in reducing the dimension of supervised structures and visualization performance. This paper is organized as follows. Section 2 deals with the related work of dimensionality reduction for classification of microarray data. Section 3 describes the dataset used, methodology and algorithms. Section 4 deals with the discussion and results. Section 5 concludes the work.

2. Related Works

This Author Jian, Linh, and David [9] presented a dimension reduction models using PLS, SIR and PCA, the comparative performance of their classification procedures were similar to PCA and PLS, the complexity of microarray data analysis was reduced. PLS and SIR were both expensive in dimension reduction but extra effective than PCA, and the results are reliable with the scrutiny of the method.

In 2012 [10] carried out normalization to regulate all the features in the dataset and dimensionality reduction to carry out clustering. The diabetic dataset which contains 768 instances and 8 attributes has been taken and PCA algorithm is used to reduce the dimensions. Out of 8 features, 4 features are selected without the loss of information.WEKA3.7 tool is used to investigate the diabetes data. After performing dimensionality reduction density based clustering algorithm is used to find the maximal set of density. Dimensionality reduction is used to increase the accuracy of the clustered data.

In 2008 [11] has compared two different feature extraction algorithms. The features of the products based on the review of the customer, is considered as the dataset. In the first algorithm the candidate features are identified and they are pruned. In the second approach association rule mining is used to find the frequent pattern. Here, the dataset is based on the customer review which are collected from the social website such as amazon,cnet and it is based on five different products(two digital cameras, a DVD player, an MP3 player and a cell phone). Likelihood Ratio Test is the method used to extract the features of the product.

In 2016 [12], has presented the methods for visual data mining in order to mine the data and to make cognitive. Here, the author has performed the attribute selection method i.e. wrapper method and filter method. Here, seven types of dataset (lung cancer, promoter, sonar, Arrhythmia, Colon Tumour, and Central Nervous System) have been used and the accuracy has been calculated before reduction and after reduction. In this reduction framework, the numbers of attributes have been reduced. The data visualization is represented in order to determine the relationship between the data. The algorithms like (LDA, QDA, and KNN) have been used in this work and found that LDA have performed efficiently and reduces the attributes effectively.

In 2014 [16] reviewed numerous development application to help users implement feature extraction of gene expression data, the paper presented review of software for feature extraction methods such as PCA, ICA, PLA and LLE. The software applications have limitations in terms of computational performance and there is need for development of classification methods to improve performances of these feature extraction methods.

2015 [16] compared dimension reduction based on logic regression models for the case-control genome-wide association by employing PCA and PLS, there were limitations in the interaction of the genes of dataset used affecting the goodness of fit and accuracy of the parameter estimation of PLS and needed further investigations.

3. Datasets Used

Colon cancer dataset was used for this experiment, it contains an expression of 2000 genes with highest minimal intensity across 62 tissues, derived from 40 tumour and 22 normal colon tissue samples [13].The gene expression was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. The gene intensity has been derived from about 20 feature pairs that correspond to the gene on the DNA microarray chip by using a filtering process. Details for data collection methods and procedures are described in [13], and the data set is available from the website http://microarray.princeton.edu/oncology/.

3.1. Methods and Algorithms

Dimensionality reduction is an important factor used in reducing original data features without the loss of information. The main objective of this analysis work is to compare two different feature extraction algorithms namely PCA and PLS. The architecture is as follows:

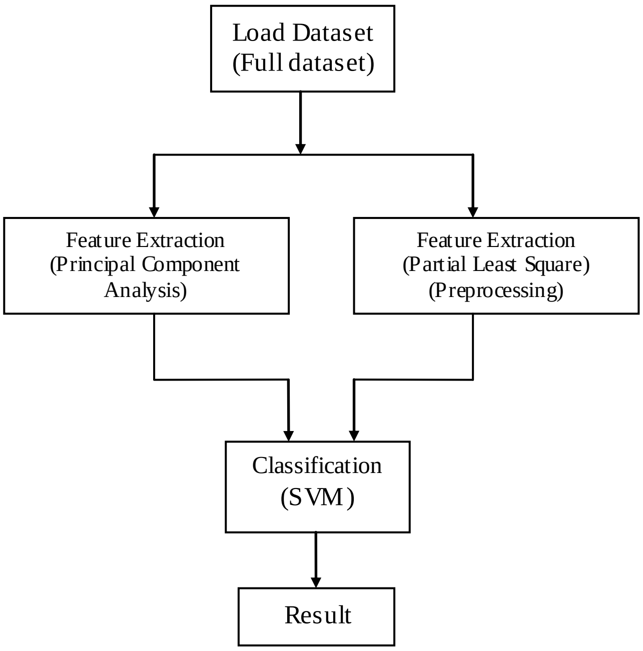


Figure 1. Technique Workflow

3.1.1. Principal Component Analysis (PCA)

In dimension reduction, PCA is one of the well-known techniques; its conception is to lessen the high dimensionality of a given dataset, while keeping enough of the variation existing in the initial predictor variables.
This is attained by transforming the *p* initial variables $X=[x_1,x_2,....,x_p]$, [16] to a latest set of *q* predictor variables.
PCA is a widely used unsupervised feature extraction technique; it works by replacing the original variables in a data with numerical variables called principal component by capturing the most descriptive features with respect to the most relevant ones [15]. PCA mathematically transforms data by referring them to a different coordinate system in order to obtain the greatest variance. A number of correlated variables into a smaller number of uncorrelated variables called principal components [10].
PCA identifies patterns of similarities and differences in a data, these patterns are determined and can be compressed by reducing the numbers of dimensions without much loss of information.
In order to conduct the PCA analysis for the input data, the following steps are performed by adopting [16]:

- Create N x d data matrix with one row vector x_n per data input.
- Subtract mean from each row vector x_n in **X**
- Calculate the covariance matrix of X
- Find Eigen Vectors and Eigen values of ∑
- Fetch the Eigen vector with the largest Eigen values.

The PCA are uncorrelated and the components explain the largest percentage in the dimensional dataset with results in extracting 10 components which are considered relevant in the colon cancer dataset used.

3.1.2. Partial Least Square (PLS)

Partial Least Square (PLS) is a supervised feature extraction technique, which is widely used as a procedure in modeling associations linking blocks of experimental variables by means of latent variable, it tries finding uncorrelated linear transformations (latent components) of the original predictor variables which have high covariance with the response variables [16]. The goal of PLS is to find the linear relationship between the response and explanatory variables y and X:

$$X = TP^T + E_x \qquad (4)$$

$$y = TC^T + E_y \qquad (5)$$

Where T represents the scores (latent variables) P and C are loadings, and E*x* and E*y* are the residual matrices obtained the original X and *y* variables.
Feature extraction using PCA ignores the response variable and its equivalence. PLS integrates the response variable during the dimensionality reduction procedure. PLS outperforms PCA in the case of microarray gene expression, PLS only consists of indicating the amount of gene components whereas PCA necessitates choosing the essential gene components [17].

3.1.3 Support Vector Machine (SVM)

In this step, the results for classification are computed using SVM for classification. SVM is a recently developed technique used for classification suggested by Vapnik, which was consecutively applied to several domains. SVM is applied to microarray cancer data which comprises of several gene expressions. SVM is applied after many steps after analysis to finally classify cancer tissues as part of an integrated algorithm. SVM is a constructive learning procedure based on statistical knowledge theory [18], it is used for classification tasks, and it uses linear models in implementing non-linear class boundaries by transforming input space using a non-linear mapping into a new space. SVM produces an accurate classifier with less over fitting and it is robust to noise.
Assuming $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ be a training set with $x_{1i} \in R^d$ and y_i is the corresponding target class. SVM can be reformulated as:
Maximize:

$$\mathbf{J} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T, x_j) \qquad (6)$$

Subject to;

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \;\; and\, \alpha_i \geq 0, i = 1, 2, \ldots, n \qquad (7)$$

This is the weighted average of the training features. Here, αi is a Lagrange multiplier of the optimization task and αi is a rank label. Values of α'si are non zero for all the points lying inside the margin and on the correct side of the classifier. The kernel function is used to solve the problem. The Kernel function analyses the relationship among the data and it creates a complex divisions in the space [19].

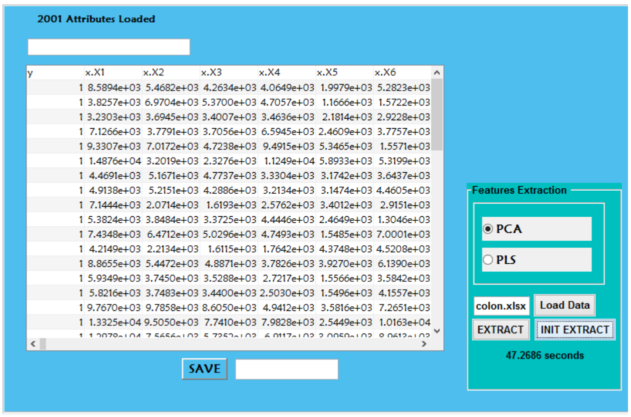
3.1.4 MATLAB

PLS, PCA, clustering, dimension reduction, factor analysis, visualization, and others. In the statistical toolbox of MATLAB, several PLS and PCA functions are provided for multivariate analysis. Most of these functions are used for dimensional reduction. All of these functions are implemented in MATLAB.

4. Results and Discussions

The colon cancer dataset extracted were classified, the classification results obtained show the features capability for classifying the colon's status. The average classification accuracy, which is using features with PCA and PLS are recorded in tabular form below.

The proposed methodology was applied to the publicly available colon cancer database.



2001 Attributes Loaded

y	x.X1	x.X2	x.X3	x.X4	x.X5	x.X6
1	8.5894e+03	5.4682e+03	4.2634e+03	4.0649e+03	1.9979e+03	5.2823e+03
1	3.8257e+03	6.9704e+03	5.3700e+03	4.7057e+03	1.1666e+03	1.5722e+03
1	3.2303e+03	3.6945e+03	3.4007e+03	3.4636e+03	2.1814e+03	2.9228e+03
1	7.1266e+03	3.7791e+03	3.7056e+03	6.5945e+03	2.4609e+03	3.7757e+03
1	9.3307e+03	7.0172e+03	4.7238e+03	9.4915e+03	5.3465e+03	1.5571e+03
1	1.4876e+04	3.2019e+03	2.3276e+03	1.1249e+04	5.8933e+03	5.3199e+03
1	4.4691e+03	5.1671e+03	4.7737e+03	3.3304e+03	3.1742e+03	3.6437e+03
1	4.9138e+03	5.2151e+03	4.2886e+03	3.2134e+03	3.1474e+03	4.4605e+03
1	7.1444e+03	2.0714e+03	1.6193e+03	2.5762e+03	3.4012e+03	2.9151e+03
1	5.3824e+03	3.8484e+03	3.3725e+03	4.4446e+03	2.4649e+03	1.3046e+03
1	7.4348e+03	6.4712e+03	5.0296e+03	4.7493e+03	1.5485e+03	7.0001e+03
1	4.2149e+03	2.2134e+03	1.6115e+03	1.7642e+03	4.3748e+03	4.5208e+03
1	8.8655e+03	5.4472e+03	4.8871e+03	3.7826e+03	3.9270e+03	6.1390e+03
1	5.9349e+03	3.7450e+03	3.5288e+03	2.7217e+03	1.5566e+03	3.5842e+03
1	5.8216e+03	3.7483e+03	3.4400e+03	2.5030e+03	1.5496e+03	4.1557e+03
1	9.7670e+03	9.7858e+03	8.6050e+03	4.9412e+03	3.5816e+03	7.2651e+03
1	1.3325e+04	9.5050e+03	7.7410e+03	7.9828e+03	2.5449e+03	1.0163e+04

SAVE

Features Extraction

● PCA

○ PLS

colon.xlsx

Load Data

EXTRACT

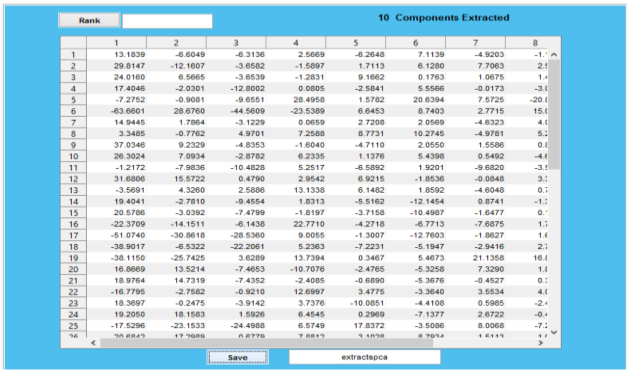
INIT EXTRACT

47.2686 seconds

Figure 1: The Loaded Colon Cancer Microarray Dataset.

Table 1: Result Evaluations

Dataset	Feature Extracted (PCA)	Feature Extracted (PLS)
Colon Cancer (2001x62)	10 Components	20 Components



Rank

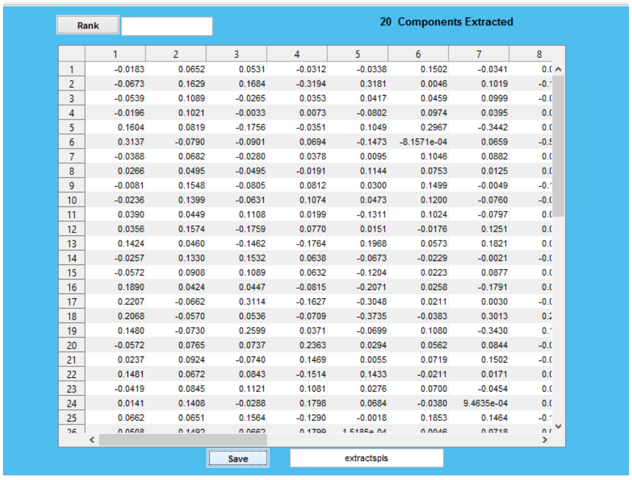
10 Components Extracted

	1	2	3	4	5	6	7
1	13.1839	-6.6049	-6.3136	2.5669	-6.2648	7.1139	-4.9203
2	29.8147	-12.1607	-3.6582	-1.5897	1.7113	6.1280	7.7063
3	24.0160	6.5665	-3.6539	-1.2831	9.1662	0.1763	1.0675
4	17.4046	-2.0301	-12.8002	0.0805	-2.5841	5.5566	-0.0173
5	-7.2752	-0.9081	-9.6551	28.4958	1.5782	20.6394	7.5725
6	-63.6601	28.6760	-44.5609	-23.5389	6.6453	8.7403	2.7715
7	14.9445	1.7864	-3.1229	0.0659	2.7208	2.0569	-4.6323
8	3.3485	-0.7762	4.9701	7.2588	8.7731	10.2745	-4.9781
9	37.0346	9.2329	-4.8353	-1.6040	-4.7110	2.0550	1.5586
10	26.3024	7.0934	-2.8782	6.2335	1.1376	5.4398	0.5492
11	-1.2172	-7.9836	-10.4828	5.2517	-6.5892	1.9201	-9.6820
12	31.6806	15.5722	0.4790	2.9542	6.9215	-1.8536	-0.0848
13	-3.5691	4.3260	2.5886	13.1338	6.1482	1.8592	-4.6048
14	19.4041	-2.7810	-9.4554	1.8313	-5.5162	-12.1454	0.8741
15	20.5786	-3.0392	-7.4799	-1.8197	-3.7158	-10.4987	-1.6477
16	-22.3709	-14.1511	-6.1438	22.7710	-4.2718	-6.7713	-7.6875
17	-51.0740	-30.8618	-28.5360	9.0055	-1.3007	-12.7603	-1.8627
18	-38.9017	-6.5322	-22.2061	5.2363	-7.2231	-5.1947	-2.9416
19	-38.1150	-25.7425	3.6289	13.7394	0.3467	5.4673	21.1358
20	16.8669	13.5214	-7.4653	-10.7076	-2.4765	-5.3258	7.3290
21	18.9764	14.7319	-7.4352	-2.4085	-0.6890	-5.3676	-0.4527
22	-16.7795	-2.7582	-0.9210	12.6997	3.4775	-3.3640	3.5534
23	18.3697	-0.2475	-3.9142	3.7376	-10.0851	-4.4108	0.5985
24	19.2050	18.1583	1.5926	6.4545	0.2969	-7.1377	2.6722
25	-17.5296	-23.1533	-24.4988	6.5749	17.8372	-3.5086	8.0068

Save

extractspca

Figure 2: PCA Features Extracted 10 Components



Rank

20 Components Extracted

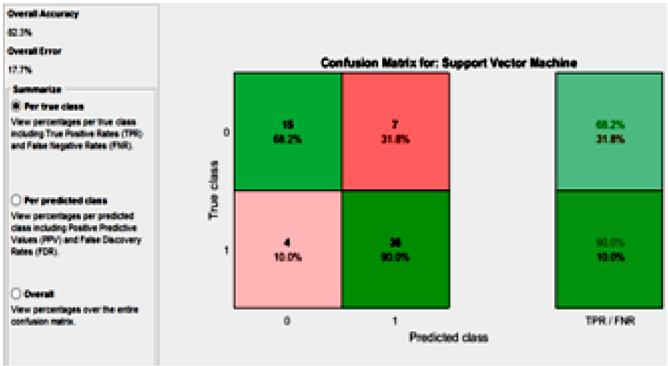
	1	2	3	4	5	6	7
1	-0.0183	0.0652	0.0531	-0.0312	-0.0338	0.1502	-0.0341
2	-0.0673	0.1629	0.1684	-0.3194	0.3181	0.0046	0.1019
3	-0.0539	0.1089	-0.0265	0.0353	0.0417	0.0459	0.0999
4	-0.0196	0.1021	-0.0033	0.0073	-0.0802	0.0974	0.0395
5	0.1604	0.0819	-0.1756	-0.0351	0.1049	0.2967	-0.3442
6	0.3137	-0.0790	-0.0901	0.0694	-0.1473	-8.1571e-04	0.0659
7	-0.0388	0.0682	-0.0280	0.0378	0.0095	0.1046	0.0882
8	0.0266	0.0495	-0.0495	-0.0191	0.1144	0.0753	0.0125
9	-0.0081	0.1548	-0.0805	0.0812	0.0300	0.1499	-0.0049
10	-0.0236	0.1399	-0.0631	0.1074	0.0473	0.1200	-0.0760
11	0.0390	0.0449	0.1108	0.0199	-0.1311	0.1024	-0.0797
12	0.0356	0.1574	-0.1759	0.0770	0.0151	-0.0176	0.1251
13	0.1424	0.0460	-0.1462	-0.1764	0.1968	0.0573	0.1821
14	-0.0257	0.1330	0.1532	0.0638	-0.0673	-0.0229	-0.0021
15	-0.0572	0.0908	0.1089	0.0632	-0.1204	0.0223	0.0877
16	0.1890	0.0424	0.0447	-0.0815	-0.2071	0.0258	-0.1791
17	0.2207	-0.0662	0.3114	-0.1627	-0.3048	0.0211	0.0030
18	0.2068	-0.0570	0.0536	-0.0709	-0.3735	-0.0383	0.3013
19	0.1480	-0.0730	0.2599	0.0371	-0.0699	0.1080	-0.3430
20	-0.0572	0.0765	0.0737	0.2363	0.0294	0.0562	0.0844
21	0.0237	0.0924	-0.0740	0.1469	0.0055	0.0719	0.1502
22	0.1481	0.0672	0.0843	-0.1514	0.1433	-0.0211	0.0171
23	-0.0419	0.0845	0.1121	0.1081	0.0276	0.0700	-0.0454
24	0.0141	0.1408	-0.0288	0.1798	0.0684	-0.0380	9.4635e-04
25	0.0662	0.0651	0.1564	-0.1290	-0.0018	0.1853	0.1464

Save

extractspls

Figure 3: PLS Feature Extracted.

In this experiment, PCA as a feature extraction method is used to reduce the high-dimension and SVM is used as the classifier. PCA is used to de-correlate the data and 10 components was achieved, in Fig. 2, the overall accuracy on all the datasets obtained using PCA as feature extraction to transform and extract the dataset is reported in a confusion matrix.



Overall Accuracy

82.3%

Overall Error

17.7%

Summarize

● Per true class

View percentages per true class including True Positive Rates (TPR) and False Negative Rates (FNR).

○ Per predicted class

View percentages per predicted class including Positive Predictive Values (PPV) and False Discovery Rates (FDR).

○ Overall

View percentages over the entire confusion matrix.

Figure 4. Confusion Matrix of Proposed Classification, using PCA-Based for Classification

True Positive Rate 68.2% and False Negative Rate yields 90.0%.

TP=36 FP=7 FN=4 TN=15

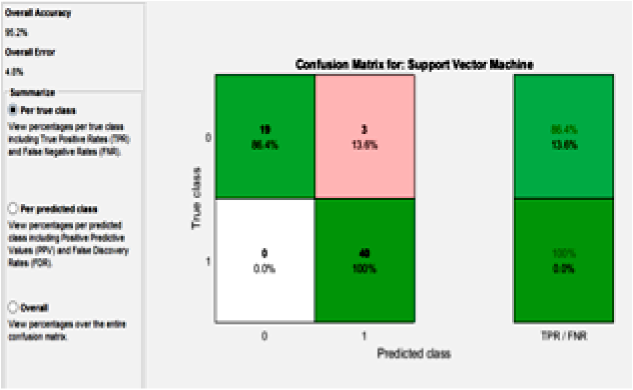
ACCURACY: (TP + TN) / (TP+TN+FP+FN) = 82.3%

SENSITIVITY: TP/ (TP+FN) = 90.00

SPECIFICITY: TN/ (FP+TN) = 68.18

PRECISION: TP/(TP+FP) = 83.72

In this experiment, to evaluate the performance of colon cancer dataset PLS is used as a feature extraction method with SVM as a classifier. It obtained a better overall accuracy compared to the first and second experiment as shown in the confusion matrix of Figure. 3, it de-correlated the features of colon cancer dataset into 20-components which are considered relevant.



Overall Accuracy
95.2%
Overall Error
4.8%

Summarize

● Per true class
View percentages per true class including True Positive Rates (TPR) and False Negative Rates (FNR).

○ Per predicted class
View percentages per predicted class including Positive Predictive Values (PPV) and False Discovery Rates (FDR).

○ Overall
View percentages over the entire confusion matrix.

Figure 5. Confusion Matrix of Proposed Classification, using PLS for Classification

True Positive Rate 86.4% and False Negative Rate yields 100.0%.

TP=40 FP=3 FN=0 TN=19
ACCURACY: (TP + TN) / (TP+TN+FP+FN) = 95.16%
SENSITIVITY: TP/ (TP+FN) = 100
SPECIFICITY: TN/ (FP+TN) = 86.36
PRECISION: TP/ (TP+FP) = 93.02

Table II: Performance Evaluation of Proposed PCA and PLS Methods.

S/No	Performance Metrics	PCA Method	PLS Method
1	Training Time	47.2686	5.0088
2	Accuracy (%)	82.30	95.16
3	Sensitivity (%)	90.00	100
4	Specificity (%)	68.18	86.36
5	Precision (%)	83.72	93.02
6	Area Under Curve	0.848864	0.994318
7	Error (%)	17.7	4.8

Table II illustrates a comparative chart between the three methods used in terms of several performance measures such as accuracy, sensitivity, specificity, precision, error, time and area under curve. This comparison shows the integrity of the proposed approach with respect to the state of the art. The colon cancer dataset used to generate our result achieved its best on PLS for feature extraction; it makes this method suitable for practitioners.

Conclusion

In this paper, a widely used colon cancer datasets was used for the evaluation of the algorithms used. The dimension reduction algorithms used to eliminate high dimensional data were PCA and PLS, it uses SVM as its classifier, and it was successfully implemented on MATLAB. For the purpose of finding the smallest gene subsets for accurate cancer classification, PLS method is highly effective compared to PCA. PLS Based method showed a better performance than PCA-based method with 95.16% to 82.30% accuracy. Hence it can be stated that the PLS based dimensionality reduction scheme is suitable for microarray gene classification as it extracts relevant and a reduced amount of information from the feature selection based technique. In future studies PLS can be compared with another feature extraction method with the aforementioned criteria. Another dataset will be a good avenue for further research of dimensionality reduction.

References

[1] Anaissi Ali, Paul Kennedy, and Goyal Mahdu: A. Dimension Reduction of Microarray Data Based on Local Principal Component, World Academy of Science, Engineering and Technology. IJCEACIE, 5(5): 529-534.

[2] Tenenbaum Joshua, Vin de Silva, and Langford John. A Global Geometric framework For Nonlinear Dimensionality reduction. Science. 290: 2319-2323 (2009).

[3] Araujo Daniel, Neto Adriao, Martins Alan, and Melo Jorge: Comparative Study on Dimension Reduction Techniques for Cluster Analysis of Microarray Data, Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA. 1(4): 1835-1842 (2011)

[4] Hua Jianping, Chao Sima, and Doughertya, Edward: Performance of feature selection methods in the classification of high-dimension data, Pattern Recognition, Volume. 42(3):409– 424 (2009).

[5] Lee John, and Verleysen Michel: Nonlinear Dimensionality Reduction, Springer Publishing Company. 8226: 617-622 (2007).

[6] Quansheng Jiang, Jiayun Lv, and Minping Jia: New approach of intelligent fault diagnosis based on LLE algorithm. Control and Decision Conference. 10: 522-526 (2008).

[7] Varini C, Nattkemper T: Breast MRI Data Analysis by LLE. Neural Networks. Proceedings. IEEE International Joint Conference. (10)3: 2449-2453 (2004).

[8] Zhang Zhenyue and Zha Hongyuan: Principal Manifolds and Nonlinear Dimensionality Reduction Via Local tangent Space Alignment. SIAM Journal of Scientific Computing. 26 (1): 313-338 (2006)

[9] Dai J, Lieu L, and Rocke D: Dimension reduction for classification with gene expression microarray data, Statistic. Appl. Genet. Mol. Biol. 5(6): 1-19 (2006)

[10] Rekha Awasthi, Anil Kumar, and Seema Pathak: An Analysis of Density Based Clustering Technique With Dimensionality Reduction for Diabetic Patient, International Journal of Computer Engineering and Applications. 9(4): 165-171 (2012)

[11] Liliana Ferreira, Niklas Jakob, and Iryna Gurevych. Comparative Study of Feature Extraction Algorithms in Customer. Proceeding of the 2nd IEEE International Conference on Semantics Computing (ICSC). 10(40): 4-7 (2008)

[12] Vijayarani S, Maria Sylviaa. Comparative Analysis of Dimensionality Reduction Techniques. International Journal of Innovative Research in Computer and Communication Engineering. 4(1): 23-29

[13] Alon U, Barkai N, Notterman D, Gish A, Ybarra S, Mack, D, and Levine, A: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl myAcad Sci USA 8; 96(12): 6745-6750 (1999)

[14] Smitarani Satpathy, and Pratikshya Mahapatra: Microarray Classification Using Intelligent Techniques, International Journal of Scientific & Engineering Research. 5(7): 1663-1670 (2014).

[15] Nebu Varghese, Vinay Verghese, Prof Gayathri. and Dr. Jaisankar: A Survey of Dimensionality Reduction and Classification Methods, International Journal of Computer Science & Engineering Survey. 3(3): 45-54 (2012).

[16] Ching Siang, Wai Soon, Mohd Saberi, Weng Howe, Safaai Deris, and Zuraini Ali: A Review of Feature Extraction Software for Microarray Gene Expression Data. BioMed Research Int. 10(1155): 15: 1-15 (2014)

[17] Honggang Yi, Hongmei Wo, Yang Zhao, Ruyang Zhang, Junchen Dai, Guangfu Jin, Hongxia Ma, Tanchun Wu, Zhibin Hu, Dongxin Lin, Hongbing Shen, and Feng Chen: Comparison of Dimension Reduction-based Logistic Regression Models for Case-control Genome-wide Association Study. The Journal of Biomedical Research. 29(4):289-307 (2015).

[18] Haozhe Xie, Jie Li, Qiaosheng Zhang, and Yadong Wang: Comparison among Dimensionality Reduction Techniques Based on Random Projection for Cancer Classification. airXiv. 2: 1-11 (2017).

[19] Seung Jun, Yichao Wu, Hao Helen, and Yufeng Liu: principal Weighted Support Vector Machine for Sufficient Dimension Reduction in Classification. Biometrika. 104(1): 67-81. (2017).