Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods

Sampath Deegalla[1] and Henrik Boström[2]

[1] Dept. of Computer and Systems Sciences,
Stockholm University and Royal Institute of Technology,
Forum 100, SE-164 40 Kista, Sweden
si-sap@dsv.su.se
[2] School of Humanities and Informatics,
University of Skövde,
P.O. Box 408, SE-541 28, Skövde, Sweden
henrik.bostrom@his.se

Abstract. Dimensionality reduction can often improve the performance of the k-nearest neighbor classifier (kNN) for high-dimensional data sets, such as microarrays. The effect of the choice of dimensionality reduction method on the predictive performance of kNN for classifying microarray data is an open issue, and four common dimensionality reduction methods, Principal Component Analysis (PCA), Random Projection (RP), Partial Least Squares (PLS) and Information Gain(IG), are compared on eight microarray data sets. It is observed that all dimensionality reduction methods result in more accurate classifiers than what is obtained from using the raw attributes. Furthermore, it is observed that both PCA and PLS reach their best accuracies with fewer components than the other two methods, and that RP needs far more components than the others to outperform kNN on the non-reduced dataset. None of the dimensionality reduction methods can be concluded to generally outperform the others, although PLS is shown to be superior on all four binary classification tasks, but the main conclusion from the study is that the choice of dimensionality reduction method can be of major importance when classifying microarrays using kNN.

1 Introduction

Microarray gene-expression technology has spread across the research community with immense speed during the last decade [1]. Being able to effectively learn from data generated through this technology is important for many reasons, including allowing for early accurate diagnoses which might lead to proper choice of treatments and therapies [2,3]. On the other hand, this type of high-dimensional data, often involving thousands of attributes, creates challenges for many learning algorithms, including the well-known k-nearest neighbor classifier (kNN) [4].

H. Yin et al. (Eds.): IDEAL 2007, LNCS 4881, pp. 800–809, 2007.

The kNN has a very simple strategy as a learner: instead of generating an explicit model, it keeps all training instances. A classification is made by measuring the distances from the test instance to all training instances, most commonly using the Euclidean distance. Finally, the majority class among the k nearest instances is assigned to the test instance. This simple form of kNN can however be both inefficient and ineffective for high-dimensional data sets due to presence of irrelevant and redundant attributes. Therefore the classification accuracy of kNN often decreases with an increase in dimensionality. One possible remedy to this problem that earlier has shown to be successful is to use dimensionality reduction [5].

The kNN has earlier been demonstrated to allow for successful classification of microarrays [2] and it has also been shown that dimensionality reduction can further improve the performance of kNN for this task [5]. However, it is an open question if the choice of dimensionality reduction technique has any impact in the performance, and for this purpose, four commonly employed dimensionality reduction methods are compared in this study when used in conjunction with kNN for microarray classification.

The organization of the paper is as follows. In the next section, we briefly present the four dimensionality reduction methods used in the study. In section 3, details of the experimental setup are provided, and the results of the comparison on eight microarray data sets are given. Finally, we give some concluding remarks and outline directions for future work.

2 Dimensionality Reduction

2.1 Principal Component Analysis (PCA)

PCA uses a linear transformation to obtain a simplified data set retaining the characteristics of the original data set.

Assume that the original matrix contains d dimensions and n observations and that one wants to reduce the matrix into a k dimensional subspace. This transformation can be given by [6]:

$$Y = E^T X \tag{1}$$

where $E_{d\times k}$ is the projection matrix containing k eigen vectors corresponding to the k highest eigen values, and $X_{d\times n}$ is the mean centered data matrix.

2.2 Random Projection (RP)

By RP, the original data set is transformed into a lower dimensional subspace by using a random matrix [7,8].

Assume that one wants to reduce the d dimensional data set into a k dimensional set where the number of instances are n. The transformation is then given by:

$$Y = R\,X \tag{2}$$

where $R_{k\times d}$ is the random matrix and $X_{d\times n}$ is the original data matrix. The original idea behind the RP is based on the Johnson-Lindenstrauss lemma (JL)

[9] which states that n points can be projected from $R^d \rightarrow R^k$ while preserving the Euclidean distance between the points within an arbitrarily small factor. For more details on the method, see [8].

This random matrix can be created in several ways and the one we have used is introduced by Achlioptas [10], by which the random matrix is generated as follows.

$$r_{ij} = \begin{cases} +\sqrt{3} \text{ with } P_r = \frac{1}{6}; \\ 0 \text{ with } P_r = \frac{2}{3}; \\ -\sqrt{3} \text{ with } P_r = \frac{1}{6}. \end{cases} \tag{3}$$

2.3 Partial Least Squares (PLS)

PLS was originally developed within the social sciences and has later been used extensively in chemometrics as a regression method [11]. It seeks for a linear combination of attributes whose correlation with the class attribute is maximized. In PLS regression the task is to build a linear model, $\bar{Y} = BX + E$, where B is the matrix of regression coefficients and E is the matrix of error coefficients. In PLS, this is done via the factor score matrix $Y = WX$ with an appropriate weight matrix W. Then it considers the linear model, $\bar{Y} = QY + E$, where Q is the matrix of regression coefficients for Y. Computation of Q will yield $\bar{Y} = BX + E$, where $B = WQ$. However, we are interested in dimensionality reduction using PLS and used the SIMPLS algorithm [12,13]. In SIMPLS, the weights are calculated by maximizing the covariance of the score vectors y_a and $\bar{y_a}$ where $a = 1, \ldots, A$ (where A is the selected numbers of PLS components) under some conditions. For more details of the method and its use, see [12,14]

2.4 Information Gain (IG)

Information Gain (IG) can be used to measure the information content in a feature [15], and is commonly used for decision tree induction. Maximizing IG is equivalent to minimizing:

$$\sum_{i=1}^{V} \frac{n_i}{N} \sum_{j=1}^{K} -\frac{n_{ij}}{n_i} \log_2 \frac{n_{ij}}{n_i}$$

where K is the number of classes, V is the number of values of the attribute, N is the total number of examples, n_i is the number of examples having the ith value of the attribute and n_{ij} is the number of examples in the latter group belonging to the jth class.

3 Empirical Study

3.1 Data Sets

The following eight microrarray data sets are used in this study:

– Colon Tumor [16], which consists of 40 tumor and 22 normal colon samples.

- Leukemia [17], which contains 72 samples of two types of leukemia: 25 acute myeloid leukemia (AML) and 47 acute lymphoblastic leukemia (ALL).
- Central Nervous System [18], which consists of 60 patient samples of survivors (39) and failures (21) after treatment of the medulloblastomas tumor (This is data set C from [18]).
- SRBCT [3], which contains four diagnostic categories of small, round blue-cell tumors as neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS).
- Lymphoma [19], which contains 42 samples of diffuse large B-cell lymphoma (DLBCL), 9 follicular lymphoma (FL) and 11 chronic lymphocytic leukemia (CLL).
- Brain [18] contains 42 patient samples of five different brain tumor types: medulloblastomas (10), malignant gliomas (10), AT/RTs (10), PNETs (8) and normal cerebella (4). (This is the data set A from [18].)
- NCI60 [20], which contains eight different tumor types. These are breast, central nervous system, colon, leukemia, melanoma, non-small cel lung carcinoma, ovarian and renal tumors.
- Prostate [2], which consists of 52 prostate tumor and 50 normal specimens.

The first three data sets come from Kent Ridge Bio-medical Data Set Repository[21] and the remaining five from [22]. The data sets are summarized in Table 1.

Table 1. Description of data

Data set	Attributes	Instances	# of Classes
Colon Tumor	2000	62	2
Leukemia	7129	38	2
Central Nervous	7129	60	2
SRBCT	2308	63	4
Lymphoma	4026	62	3
Brain	5597	42	5
NCI60	5244	61	8
Prostate	6033	102	2

3.2 Experimental Setup

We have used Matlab to transform raw attributes to both PLS and PCA components. The PCA transformation is performed using the Matlab's Statistics Toolbox whereas the PLS transformation is performed using the BDK-SOMPLS toolbox[23,24], which uses the SIMPLS algorithm. The WEKA data mining toolkit [15] is used for the RP and IG methods, as well as for the actual nearest neighbor classification.

Both PLS and IG are supervised methods which use class information for their transformations. Therefore, to generate the PLS components for test sets, the

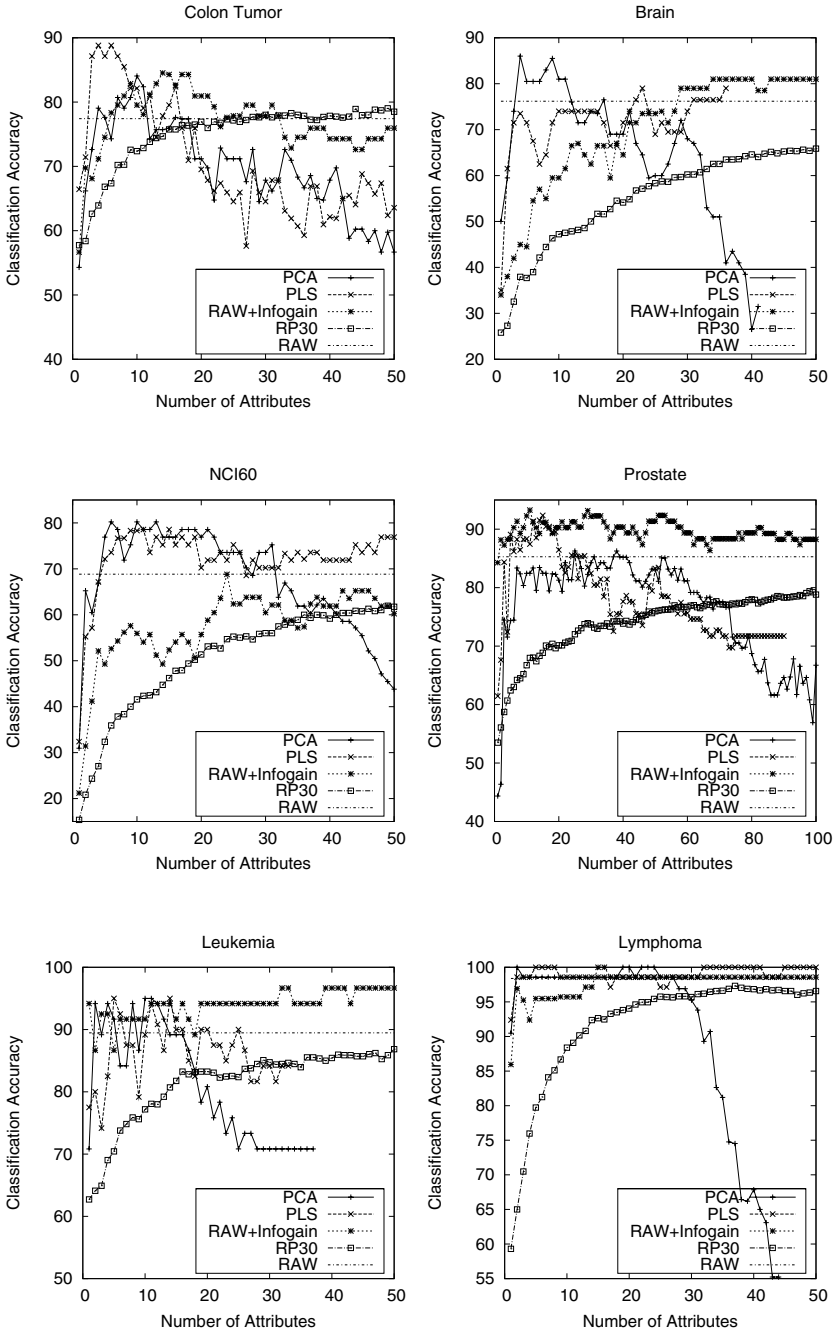

Fig. 1. Predictive performance with the change of numbers of dimensions using PCA, PLS, RP and IG with Nearest Neighbor (IB1) for Colon Tumor, Brain, NCI60, Prostate, Leukemia and Lymphoma data sets

weight matrix generated for the training set has to be used. For IG, attributes in the training set are ranked based on the information content in a decreasing manner and the same attributes are selected for the test set. As earlier explained, attributes generated using RP are of a random nature since a random matrix is used for the transformation. For this reason, we have averaged results of RP from 30 runs to reduce the variance.

The optimal number of neighbors (i.e., k) could be specific to different data sets and dimensionality reduction methods. Therefore, we have investigated the effect of different values of k, namely 1, 3, 5, 7 and 9.

Stratified 10-fold cross validation [15] is employed to obtain measures of accuracy, which has been chosen as the performance measure in this study.

3.3 Experimental Results

The results are summarized in Fig. 1 and Fig. 2. It can be observed that both PLS and PCA obtain their best classification accuracies with relatively few dimensions, while more dimensions are required for IG and many more for RP.

None of the methods turns out as a clear winner, except perhaps PLS on the binary classification tasks. However, all methods outperform not using dimensionality reduction, and the difference in performance between the best and worst method can vary greatly for a particular dataset, leading to the conclusion that the choice of dimensionality reduction to be used in conjunction with kNN for microarray classification can be of major importance.

In most of the cases, simply setting $k = 1$ gives the best result. However, for IG it seems that one should consider choosing higher values for k which improves the classification accuracy by at least 1% for 5 out of 8 datasets. For PCA, the choice of a higher k value yields at least a 1% improvement for 3 out of 8 data sets whereas for PLS, an improvement of at least 1% is obtained for 4 out of 8 datasets.

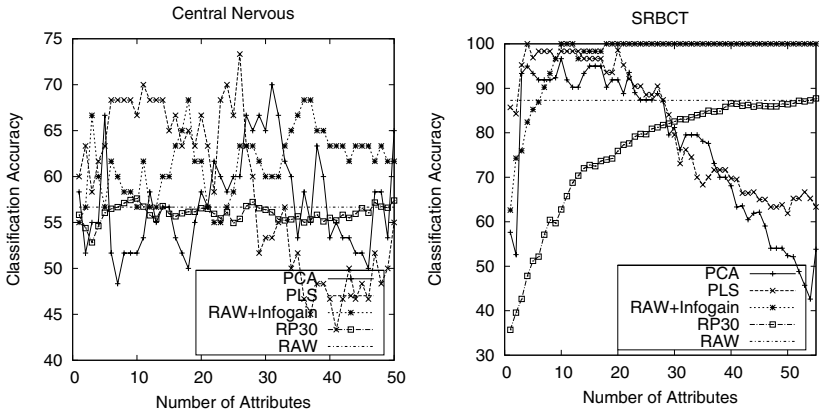

Fig. 2. Predictive performance with the change of numbers of dimensions using PCA, PLS, RP and IG with Nearest Neighbor (IB1) for Central Nervous and SRBCT data sets

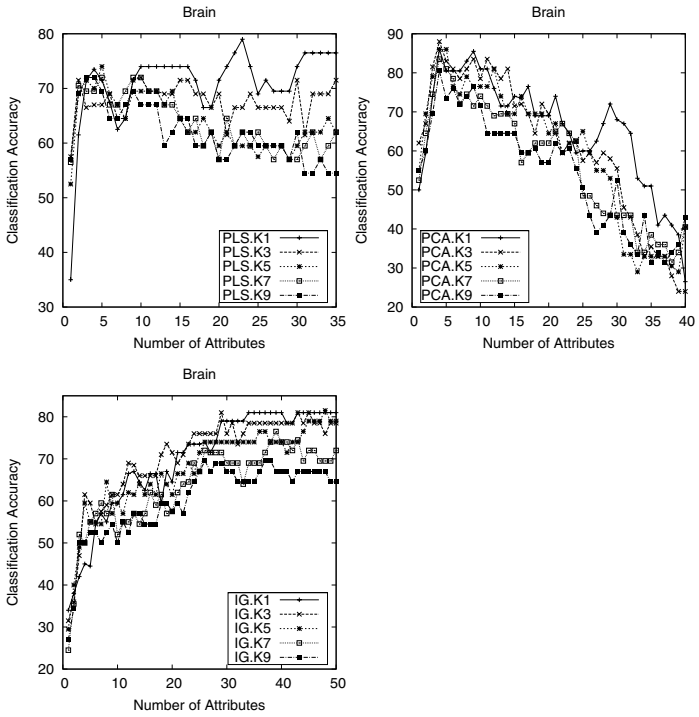

Fig. 3. Predictive accuracy with different k values for nearest neighbor classifier for Brain dataset

Table 2. Order of k values w.r.t averaged accuracy

	Decreasing order of accuracy		
	IG	PCA	PLS
ColonTumor	7,5,9,3,1	5,9,7,3,1	7,9,5,3,1
Leukemia	1,3,5,7,9	1,3,5,7,9	3,1,5,7,9
CentralNervous	7,9,5,1,3	3,7,9,5,1	9,7,5,3,1
SRBCT	3,5,1,7,9	1,9,3,7,5	9,7,5,3,1
Lymphoma	5,9,1,7,5	1,3,5,7,9	1,3,5,7,9
Brain	3,1,5,7,9	1,3,5,7,9	1,3,5,7,9
NCI60	9,7,1,5,3	1,3,5,7,9	1,3,5,7,9
Prostate	3,7,9,5,1	9,5,7,3,1	9,3,1,7,5

4 Concluding Remarks

Four dimensionality reduction methods are compared for classifying microarrays with the nearest neighbor classifier. Experiments with eight microarray datasets show that dimensionality reduction indeed is effective for nearest neighbor classification.

However, none of the methods used in the study consistently gives the best accuracy on all data sets. Generally, both PCA and PLS results in the highest accuracy for few dimensions whereas RP and IG require more dimensions. Compared to the other three methods, PCA is shown to be more sensitive to the choice of dimensionality, and typically gives poor results in higher dimensions. It can be observed that PLS outperforms the other methods for binary classification problems (Colon, Leukemia, Central Nervous and Prostate).

We have also investigated the accuracy of kNN for different values of k. Generally, k=1 seems to be the best choice for PCA and PLS, while higher values are required for IG.

There are a number of issues that need further exploration. First, additional binary microarray classification tasks could be investigated to test the finding that PLS appears to be superior in these cases. Second, further characterizations of the situations in which the different dimensionality reduction methods are successful could be identified. Furthermore, the possibility of combining several reduced features sets generated by different reduction methods could also be investigated.

Acknowledgements

Financial support from SIDA/SAREC for the first author is greatly acknowledged. The second author was supported by the Information Fusion Research Program (www.infofusion.se) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2003/0104.

References

1. Quackenbush, J.: Microarray analysis and tumor classification. The New England Journal of Medicine 354(23), 2463–2472 (2006)
2. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209 (2002)
3. Kahn, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., Meltzer, P.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7, 673–679 (2001)
4. Aha, D.W., Kiblear, D., Albert, M.K.: Instance based learning algorithm. Machine Learning 6, 37–66 (1991)
5. Deegalla, S., Bostrom, H.: Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In: ICMLA 2006. Proceedings of the 5th International Conference on Machine Learning and Applications, pp. 245–250. IEEE Computer Society, Washington, DC, USA (2006)
6. Shlens, J.: A tutorial on principal component analysis, `http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf`

7. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: KDD 2001. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 245–250 (2001)
8. Fradkin, D., Madigan, D.: Experiments with random projections for machine learning. In: KDD 2003. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 517–522 (2003)
9. Dasgupta, S., Gupta, A.: An elementary proof of the Johnson-Lindenstrauss lemma. Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, USA (1999)
10. Achlioptas, D.: Database-friendly random projections. In: ACM Symposium on the Principles of Database Systems, pp. 274–281 (2001)
11. Abdi, H.: Partial least squares (pls) regression (2003)
12. de Jong, S.: SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems (1993)
13. StatSoft Inc.: Electronic statistics textbook (2006), `http://www.statsoft.com/textbook/stathome.html`
14. Boulesteix, A.L.: Pls dimension reduction for classification with microarray data. Statistical Applications in Genetics and Molecular Biology (2004)
15. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)
16. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In: Proc. Natl. Acad. Sci., vol. 96, pp. 6745–6750 (1999)
17. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
18. Pomeroy, S.L., Tamayo, P., Gassenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415, 436–442 (2002)
19. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511 (2000)
20. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., de Rijn, M.V., Waltham, M., Pergamenschikov, A., Lee, J.C, Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O.: Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics 24(3), 227–235 (2000)
21. Kent Ridge Bio-medical Data Set Repository, `http://sdmc.lit.org.sg/GEDatasets/Datasets.html`

22. Díaz-Uriarte, R., de Andrés, S.A.: Gene selection and classification of microarray data using random forest. Bioinformatics 7(3) (2006), http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html
23. Melssen, W., Wehrens, R., Buydens, L.: Supervised kohonen networks for classification problems. Chemometrics and Intelligent Laboratory Systems 83, 99–113 (2006)
24. Melssen, W., Üstün, B., Buydens, L.: Sompls: a supervised self-organising map - partial least squares algorithm. Chemometrics and Intelligent Laboratory Systems 86(1), 102–120 (2006)