Fusion of Dimensionality Reduction Methods: a Case Study in Microarray Classification

Sampath Deegalla
Dept. of Computer and Systems Sciences
Stockholm University
Sweden
si-sap@dsv.su.se

Henrik Boström
Informatics Research Centre
University of Skövde
Sweden
henrik.bostrom@his.se

Abstract – *Dimensionality reduction has been demonstrated to improve the performance of the k-nearest neighbor (kNN) classifier for high-dimensional data sets, such as microarrays. However, the effectiveness of different dimensionality reduction methods varies, and it has been shown that no single method constantly outperforms the others. In contrast to using a single method, two approaches to fusing the result of applying dimensionality reduction methods are investigated: feature fusion and classifier fusion. It is shown that by fusing the output of multiple dimensionality reduction techniques, either by fusing the reduced features or by fusing the output of the resulting classifiers, both higher accuracy and higher robustness towards the choice of number of dimensions is obtained.*

Keywords: Nearest neighbor classification, dimensionality reduction, feature fusion, classifier fusion, microarrays.

1 Introduction

There is a strong need for accurate methods for analyzing microarray gene-expression data sets, since early accurate diagnoses based on these analyses may lead to proper choice of treatments and therapies [1, 2, 3]. However, the nature of these data sets (i.e., thousands of attributes with small number of instances) is a challenge for many learning algorithms, including the well-known k-nearest neighbor (kNN) classifier [4].

The kNN has a very simple strategy as a learner: instead of generating an explicit model, it keeps all training instances. 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, i.e., projecting the original feature set into a smaller number of features [5].

The use of 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, different dimensionality reduction methods may have different effects on the performance of the kNN classifier, and it has been shown that no single method always outperforms the others when used for microarray classification [6]. As an alternative to choosing a single method, we will in this study consider the idea of applying a set of dimensionality reduction methods and fusing the output of these. Two fusion approaches are investigated: feature fusion, i.e., combining the reduced subset of features before learning with kNN, and classifier fusion, i.e., combining the individual kNN classifiers built from each feature reduction method.

The organization of the paper is as follows. In the next section, we briefly present three dimensionality reduction methods that will be considered in the investigation together with the approaches for combining (or fusing) the output of these. 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 in section 4.

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 o dimensions and n observations and that one wants to reduce the matrix into a d dimensional subspace. Following [7], this transformation can be defined by:

$$Y = E^T X \quad (1)$$

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

2.2 Partial Least Squares (PLS)

PLS was originally developed within the social sciences and has later been used extensively in chemometrics as a regression method [8]. It seeks for a linear combination of attributes whose correlation with the class attribute is maximum.

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 [9, 10]. In SIMPLS, the weights are calculated by maximizing the covariance of the score vectors y_a and $\bar{y_a}$ where $a = 1, \ldots, d$ (where d is the selected number of PLS components) under some conditions. For more details of the method and its use, see [9, 11].

2.3 Information Gain (IG)

Information Gain (IG) can be used to measure the information content in a feature [12], and is commonly used for decision tree induction. Maximizing IG is equivalent to minimizing:

$$\sum_{i=1}^{V} \frac{n_i}{N} \sum_{j=1}^{C} -\frac{n_{ij}}{n_i} \log_2 \frac{n_{ij}}{n_i} \qquad (2)$$

where C 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.

When it comes to feature reduction with IG, all features are ranked according to decreasing information gain, and the first d features are selected.

It is also necessary to consider how discretization of numerical features is to be done. Since such features are present in all the considered data sets, they have to be converted to categorical features in order to allow for the use of the above calculation of IG. We used the WEKA's default configuration, i.e., Fayyad & Irani's Minimum Description Length (MDL) [13] method, for discretization.

2.4 Feature fusion (FF)

Feature fusion concerns how to generate and select a single set of features for a set of objects to which several sets of features are associated [14]. In this study, we use a single data source together with different dimensionality reduction methods which allows us to perform feature fusion by concatenating features generated by the different methods. High-dimensionality is not a problem here since each transformed data set is small compared to the original size of the data. Therefore, a straightforward method of choosing an equal number of features from each reduced set is considered. The selected total number of dimensions are from d = 3 to 99. For each d, the first $d/3$ reduced dimensions are chosen from the output of PLS, PCA and IG respectively.

2.5 Classifier fusion (CF)

The focus of classifier fusion is either on generating a structure representing a set of combined classifiers or on combining classifier outputs [15]. We have considered the latter approach, i.e., combining nearest neighbor predictions with PLS, PCA and IG using unweighted voting. For multi-class problems, ties are resolved by randomly selecting one of the predictions.

3 Empirical study

3.1 Data sets

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

- Central Nervous System [16], which consists of 60 patient samples of survivors (39) and failures (21) after treatment of the medulloblastomas tumor (data set C from [16]).
- Colon Tumor [17], which consists of 40 tumor and 22 normal colon samples.
- Leukemia [18], which contains 72 samples of two types of leukemia: 25 acute myeloid leukemia (AML) and 47 acute lymphoblastic leukemia (ALL).
- Prostate [2], which consists of 52 prostate tumor and 50 normal specimens.
- Brain [16] contains 42 patient samples of five different brain tumor types: medulloblastomas (10), malignant gliomas (10), AT/RTs (10), PNETs (8) and normal cerebella (4) (data set A from [16]).
- Lymphoma [19], which contains 42 samples of diffuse large B-cell lymphoma (DLBCL), 9 follicular lymphoma (FL) and 11 chronic lymphocytic leukemia (CLL).
- NCI60 [20], which contains eight different tumor types. These are breast, central nervous system, colon, leukemia, melanoma, non-small cell lung carcinoma, ovarian and renal tumors.

Table 1: Description of data

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

- 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).

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

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 [12] is used for the IG method, as well as for the nearest neighbor classification.

Both PLS and IG are supervised methods which use class information for their transformations. Therefore, to generate the PLS components for a test set, for which the class labels are unknown, the weight matrix generated for the training set has to be used. For IG, attributes in the training set is ranked based on the information gain in a decreasing manner and the same attributes are selected for the test set. The numbers of selected dimensions were varied from one to approximately the number of examples in the current data set for all three methods.

Stratified 10-fold cross validation [12] is employed to obtain measures of accuracy, which has been chosen as the performance indicator in this study.

3.3 Experimental results

The results of using the original features, the three dimensionality reduction methods, the feature fusion and classifier fusion methods are shown in Fig. 1 and Fig. 2. Table 2 summarizes the highest classification accuracies of individual reduction and fusion methods comparing their accuracy to raw features. The numbers inside the brackets denote the minimum number of dimensions required to reach a particular accuracy.

It can be observed in the Fig. 1 and Fig. 2 that both PLS and PCA obtain their best classification accuracies with relatively few dimensions, while more dimensions are required for IG. None of the single methods turns out as a clear winner, except for perhaps PLS on the binary classification tasks. However, all three methods outperform not using dimensionality reduction, and the difference in performance between the best and worst method can vary greatly for a particular data set, 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, but also that no single method is suitable for all cases.

As shown in Table 2, the feature fusion method often performs well compared to each individual method, giving the overall best results in 7 out of 8 cases. Furthermore, the accuracy varies to a much less extent with the number of dimensions compared to the other methods, hence showing that using the combined features reduces sensitivity to the choice of the number of dimensions. In addition, the classifier fusion method also performs on par with or better than the individual dimensionality reduction methods in 6 out of 8 cases, reaching the overall best accuracy in 3 cases. It should also be noted that classifier fusion yields the best accuracy with fewer dimensions compared to the feature fusion method. For example, the FF method reaches the highest classification accuracy with 60 dimensions for the Central Nervous data set, whereas the CF method reaches its highest level using only 11 features.

4 Concluding remarks

Three dimensionality reduction methods were investigated in conjunction with the kNN classifier and two approaches to fusing the result of multiple dimensionality reduction methods were considered: feature fusion and classifier fusion. An experiment with eight microarray data sets shows that dimensionality reduction indeed is effective for nearest neighbor classification and that fusing the output of these methods can further improve the classification accuracy compared to the individual dimensionality reduction methods.

It is also observed that PCA and PLS are best when choosing few dimensions and they even sometimes outperform the fusion methods. However, if one compares the best classification accuracies between individual methods and fusion methods, the feature fusion method obtain the best classification accuracy in 7 out of 8 cases. In addition, classifier fusion obtains the best accuracy in relatively few number of dimensions compared to feature fusion which on the other hand is even more robust to changes in number of dimensions. Therefore, it can be concluded that choosing any of the fusion approaches should be preferred to choosing any of the single dimensionality reduction methods, since the former can be expected to lead to higher classification accuracy and robustness with respect to the choice of number of dimensions.

Table 2: Results on best classification accuracies

Data set	Raw	PCA	PLS	IG	FF	CF
Central Nervous	56.67	70.00(31)	**73.33(26)**	68.33(18)	**73.33(60)**	71.67(11)
Colon Tumor	77.42	84.05(10)	**88.81(4)**	84.52(14)	87.14(21)	87.38(7)
Leukemia	89.47	95.00(10)	95.00(5)	96.67(32)	**100.00(33)**	97.50(4)
Prostate	85.29	86.27(25)	92.36(15)	93.27(11)	**95.18(87)**	95.09(50)
Brain	76.19	86.00(4)	79.00(23)	81.00(34)	**88.00(96)**	86.00(4)
Lymphoma	98.39	**100.00(2)**	**100.00(5)**	**100.00(15)**	**100.00(6)**	**100.00(2)**
NCI60	68.85	**80.24(6)**	78.57(11)	68.81(24)	**80.24(51)**	**80.24(6)**
SRBCT	87.30	96.67(10)	**100.00(4)**	**100.00(10)**	**100.00(12)**	**100.00(4)**

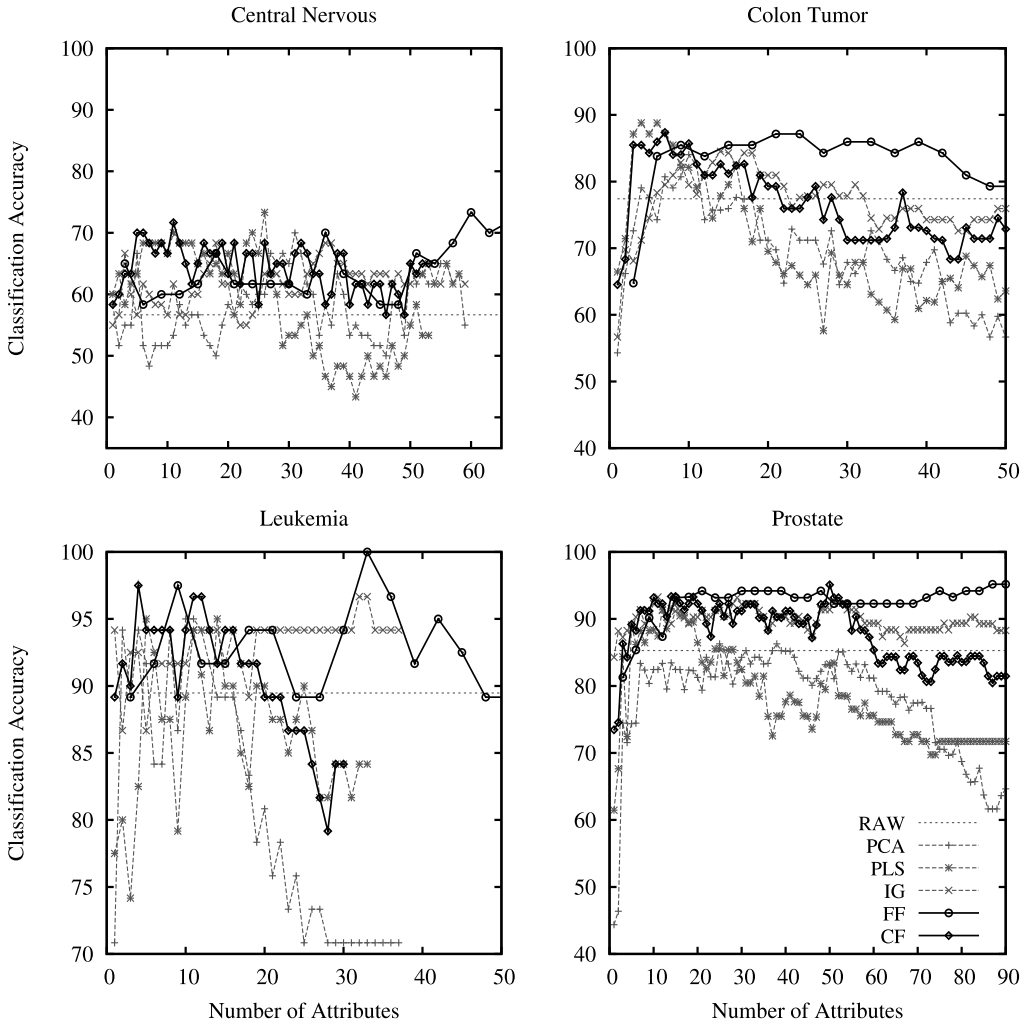

Figure 1: Predictive performance with the change of numbers of dimensions using PCA, PLS, IG, Feature fusion (FF) and Classifier fusion (CF) with Nearest Neighbor for two class microarray data sets.

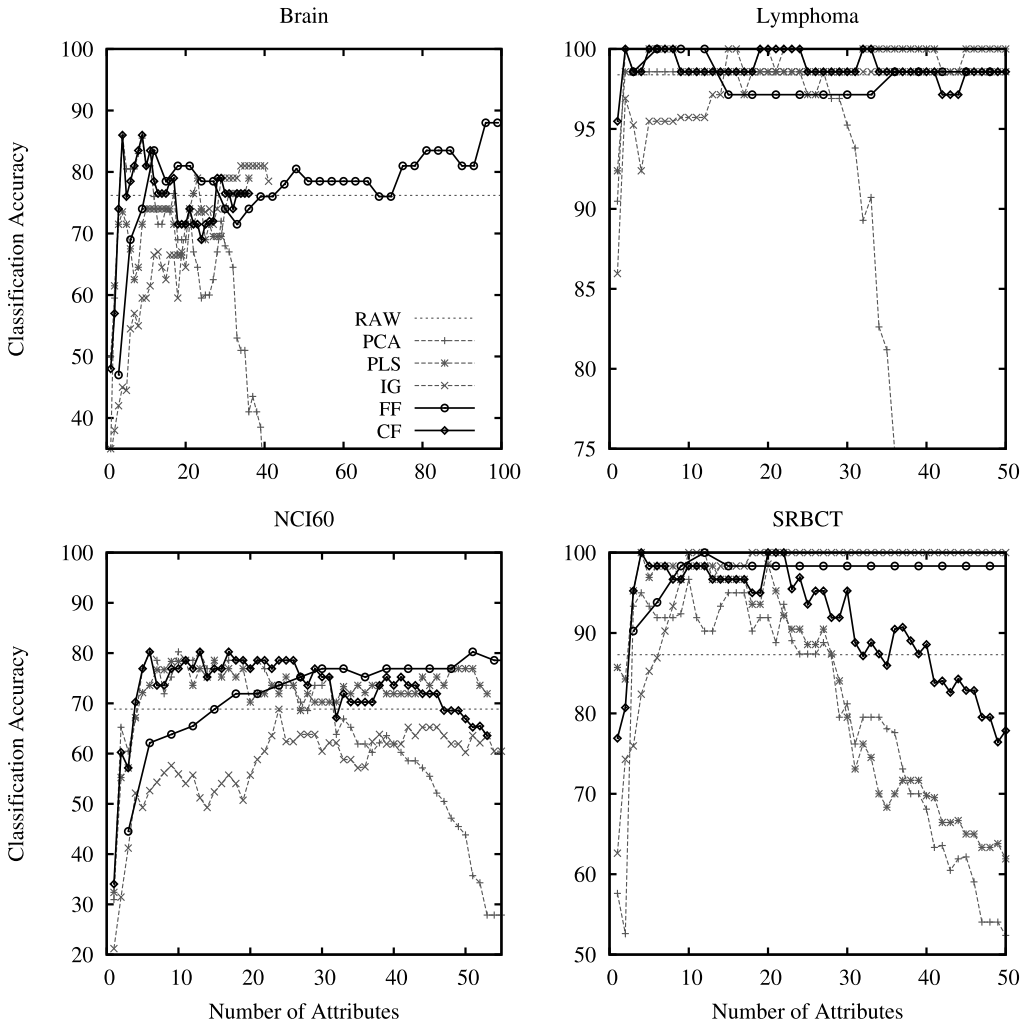

Figure 2: Predictive performance with the change of numbers of dimensions using PCA, PLS, IG, Feature fusion (FF) and Classifier fusion (CF) with Nearest Neighbor for multi class microarray data sets.

There are a number of issues that need further exploration. First, fusion of classifiers from additional dimensionality reduction methods could be investigated, e.g., Random Projection [6]. Second, selecting different number of dimensions for different dimensionality reduction methods when fusing classifiers of features, could be investigated as an alternative strategy. Finally, more sophisticated voting strategies could be considered, see e.g., [25].

5 Acknowledgments

The first author gratefully acknowledge support through the SIDA/SAREC IT project. This work 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] J. Quackenbush. Microarray analysis and tumor classification. *The New England Journal of Medicine*, 354(23):2463–2472, 2006.

[2] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

[3] J. Kahn, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.

[4] D. W. Aha, D. Kiblear, and M. K. Albert. Instance based learning algorithm. *Machine Learning*, 6:37–66, 1991.

[5] S. Deegalla and H. Boström. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *ICMLA '06: Proceedings of the 5th International Conference on Machine Learning and Applications*, pages 245–250, Washington, DC, USA, 2006. IEEE Computer Society.

[6] S. Deegalla and H. Boström. Classification of microarrays with knn: Comparison of dimensionality reduction methods. In *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning, Birmingham, UK*, pages 800–809, 2007.

[7] J. Shlens. A tutorial on principal component analysis. `http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf`.

[8] H. Abdi. Partial least squares (pls) regression., 2003.

[9] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 1993.

[10] StatSoft Inc. Electronic statistics textbook, 2006. `http://www.statsoft.com/textbook/stathome.html`.

[11] A. Boulesteix. PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2004.

[12] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.

[13] U. M. Fayyad and K. B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8:87–102, 1992.

[14] H. Boström. Feature vs. classifier fusion for predictive data mining - a case study in pesticide classification. In *Proceedings of the 10th International Conference on Information Fusion*, 2007.

[15] D. Ruta and B. Gabrys. An overview of classifier fusion methods. *Computing and Information Systems*, 7:1–10, 2000.

[16] S. L. Pomeroy, P. Tamayo, M. Gassenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, January 2002.

[17] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proc. Natl. Acad. Sci. USA*, volume 96, pages 6745–6750, 1999.

[18] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[19] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[20] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. V. de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. 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://datam.i2r.a-star.edu.sg/datasets/krbd/index.html`.

[22] R. Díaz-Uriarte and S. A. de Andrés. Gene selection and classification of microarray data using random forest. *Bioinformatics*, 7(3), 2006. `http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html`.

[23] W. Melssen, R. Wehrens, and L. Buydens. Supervised kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems*, 83:99–113, 2006.

[24] W. Melssen, B. Üstün, and L. Buydens. Sompls: a supervised self-organising map - partial least squares algorithm. *Chemometrics and Intelligent Laboratory Systems*, 86(1):102–120, 2006.

[25] H. Boström, R. Johansson, and A. Karlsson. On evidential combination rules for ensemble classifiers. In *Proceedings of the 11th International Conference on Information Fusion*, pages 553–560, 2008.