Improving Fusion of Dimensionality Reduction Methods for Nearest Neighbor Classification

Sampath Deegalla and Henrik Boström
Department of Computer and Systems Sciences
Stockholm University
Forum 100, SE-164 40 Kista, Sweden
Email: si-sap@dsv.su.se, henrik.bostrom@dsv.su.se

***Abstract*—In previous studies, performance improvement of nearest neighbor classification of high dimensional data, such as microarrays, has been investigated using dimensionality reduction. It has been demonstrated that the fusion of dimensionality reduction methods, either by fusing classifiers obtained from each set of reduced features, or by fusing all reduced features are better than than using any single dimensionality reduction method. However, none of the fusion methods consistently outperform the use of a single dimensionality reduction method. Therefore, a new way of fusing features and classifiers is proposed, which is based on searching for the optimal number of dimensions for each considered dimensionality reduction method. An empirical evaluation on microarray classification is presented, comparing classifier and feature fusion with and without the proposed method, in conjunction with three dimensionality reduction methods; Principal Component Analysis (PCA), Partial Least Squares (PLS) and Information Gain (IG). The new classifier fusion method outperforms the previous in 4 out of 8 cases, and is on par with the best single dimensionality reduction method. The novel feature fusion method is however outperformed by the previous method, which selects the same number of features from each dimensionality reduction method. Hence, it is concluded that the idea of optimizing the number of features separately for each dimensionality reduction method can only be recommended for classifier fusion.**

***Keywords*-nearest neighbor classification; dimensionality reduction; feature fusion; classifier fusion; microarrays;**

I. Introduction

Microarray gene expression technology has enhanced the accurate identification of cancer. Great precision is essential in many cases, since early identification of cancer may often lead to proper choice of treatments and therapies [1], [2], [3]. The accuracy of the k nearest neighbor classifier (kNN) is generally low for classification of microarrays due to the nature of these data sets, i.e., the number of instances is very low compared to the number of dimensions [4]. To improve the classification accuracy of kNN, one may employ dimensionality reduction to reduce the original high dimensionality into lower numbers of dimensions.

In a previous study [5], three dimensionality reduction methods, i.e., Principal Component Analysis (PCA), Partial Least Squares (PLS) and Information Gain (IG), were compared w.r.t. the classification accuracy when used in conjunction with kNN. It was shown that all dimensionality reduction methods improve the classification accuracy of kNN compared to using the original features. The results also showed that none of the single methods was able to outperform the others, and therefore, feature fusion, i.e., combining feature subsets from different dimensionality reduction methods, and classifier fusion, i.e., combining outputs of the classifiers from each dimensionality reduction were considered in a subsequent study [6]. In that study, two methods were considered: simple straight forward combination of features and combining classifier outputs using unweighted voting. It was shown that the fusion methods not only improve the performance of kNN, but also that they were robust to changes in dimensionality, i.e., the choice of number of dimensions did not affect the accuracy to a high extent. However, none of the previous fusion methods were shown to constantly improve the classification accuracy of kNN compared to a single dimensionality reduction method. One reason for this could be that the same number of dimensions were chosen for all dimensionality reduction methods, although previous studies have shown that the best number of dimensions to choose for each individual method may vary significantly. In this study, we address this problem by extending the fusion methods by the capability of searching for the optimal number of dimensions for each dimensionality reduction method.

In the next section, the three considered dimensionality reduction methods are described together with different ways to fuse features and classifiers, including the proposed method of searching for the optimal number of dimensions. An experiment comparing the fusion and dimensionality reduction on eight microarray data sets is presented in section III. Finally, concluding remarks are given in section IV.

II. Methods

A. Principal Component Analysis (PCA)

PCA is a classical dimensionality reduction method that has been applied in many different contexts, including face recognition, image compression, cancer classification and applications related to high-dimensional data sets. This

method is well known for allowing the original dimensionality to be reduced a much smaller, uncorrelated feature set with minimum information loss. Transformed features are generally known as principal components, which are weighted linear combinations of original features and which are orthogonal to each other. The components are typically ordered according to decreasing variability, i.e., the first principal component has the highest variability in the data set, the second principal component has the second highest and so on.

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:

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

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.

B. 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 features whose correlation with the output variable 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].

C. 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:

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

where C is the number of classes, V is the number of values of the feature, N is the total number of examples, n_i is the number of examples having the ith value of the feature 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 data sets considered in this study, they have to be converted to categorical features in order to allow for the use of the above calculation of IG. We used WEKA's default configuration, i.e., Fayyad and Irani's Minimum Description Length (MDL) [13] method for discretization.

D. 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 have investigated two possible feature fusion methods. The first method has been investigated in [6] and the second method is one of the two new extended fusion methods.

In the first method, which is denoted by FF1, an equal number of features from each dimensionality reduction method are considered for classification with kNN. Therefore, the total number of dimensions selected for classification is in the range $d = 3 \ldots 3 \times (n_t - 1)$[1] where n_t is the number of training instances. For each d, the first $d/3$ reduced dimensions are chosen from the output of PLS, PCA and IG respectively.

In the second method, which is denoted by FF2, the minimum number of features required to yield the highest classification accuracy for each dimensionality reduction method is considered. Cross-validation is performed to find this number. For example, to find the minimum number of features required to get the best performance for PCA, cross-validation is performed to find which number of principal components should be selected to get the highest accuracy by kNN, when projecting the initial features to this set. During the cross-validation, which uses only the provided set of training examples, part of the examples are used to generate the components, while the remaining part is used to estimate the classifier's performance.

E. 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 when used in conjunction with PLS, PCA and IG.

In the first method, denoted by CF1, for each dimension, nearest neighbor predictions from each dimensionality reduction method are combined using unweighted voting, i.e.,

[1] $(n_t - 1)$ provides an upper bound on the number of features generated by all three methods, since this is the maximum number of features generated by PLS.

giving equal weight to the output of each nearest neighbor classifier and selecting the majority output among them. This method, which has been investigated in a previous study [6], hence combines classifiers generated from different projections into a specified number of dimensions, i.e., all combined classifiers reduce the original feature set into the same number of dimensions similarly to the method FF1 above.

In the novel method, denoted by CF2, the number of dimensions is selected that results in the highest accuracy for each dimensionality reduction method, as estimated by cross-validation on the training set. The outputs are then fused using unweighted voting. Both classifier fusion methods may lead to ties for multi-class problems which are resolved by randomly selecting one of the class labels achieving the highest number of votes.

III. Empirical study

A. 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.
- 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 I.

Table I
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

B. Experimental setup

First, raw features are transformed into a lower number of dimensions using all reduction methods. PCA and PLS transformations are applied to the training set and the generated weight matrix is used to transform the test set. In IG, features based on the information content are ranked in decreasing manner in the training set and the same rankings are used when classifying the test set. For kNN, k=1 is considered, i.e., a single nearest neighbor is chosen. To find the optimal number of features, cross-validation is performed using the training set with the nearest neighbor classifier. Then the optimal number of features for the training set is selected for the final classification. Nearest neighbor classification is performed on the reduced space generated from the training set with optimal number of features by PCA, PLS and IG.

Then, different ways of fusing features and classifiers are considered. For each number of dimensions, straightforward combination of features, which are generated by the three dimensionality reduction methods, are considered by FF1. In CF1, nearest neighbor classifier outputs generated by PCA, PLS and IG are considered for each dimension. For each output of CF1, the majority class among the three classifiers generated by PCA, PLS and IG is considered. For each training set, cross-validation is performed to find the optimal number of dimensions.

In the new feature fusion approach, i.e., FF2, for each dimensionality reduction method, the optimal number of features that gives the best performance on the reduced space generated by that method are found and then fused together for classification. The same number of dimensions are then selected in the test set and combined. In the new classifier fusion method, i.e., CF2, the classifier which gives the best performance using cross-validation on the training set for each dimensionality reduction method is selected. The output of the three best classifiers on the training set are then fused together using unweighted voting, i.e., getting the majority class among the outputs.

For overall validation, 10-fold cross-validation was em-

ployed, i.e., the data set is divided into ten folds and then, in each iteration, nine folds are taken as the training set and the rest as the testing set. To find the optimal number of features in each training set, 10-fold cross-validation is used again. PCA transformation is conducted using the Matlab's Statistics Toolbox and PLS transformation is conducted using the BDK-SOMPLS toolbox [23], [24], which uses the SIMPLS algorithm. IG and nearest neighbor classification are performed using the WEKA data mining toolkit [12].

C. Experimental results

Table II lists the classification accuracies obtained using original features, the three dimensionality reduction methods, the two feature fusion methods (FF1, FF2) and the two classifier fusion methods (CF1, CF2).

As can be seen in Table II, the new methods, i.e., FF2 and CF2, outperform using the original feature set in 6 and 7 data sets respectively, out of 8. If one compares the two methods, CF2 gets the best accuracy in four cases, while FF2 gets the best accuracy in 3 cases. If one compares the novel fusion methods with each individual dimensionality reduction method, each novel method outperform PCA and IG in 7 out of 8 cases whereas for PLS there is a tie with CF2 in 5 out of 8 cases. The trend in the results are quite different from the results reported in [6], mainly due to the change of the experimental setting. We have here only considered the optimal number of features for each method whereas all dimensions are considered in [6] from which the best individual result is selected and compared to the other results. Although the results show that the use of PLS outperforms raw accuracies in all cases, this is not the case for PCA and IG, which also was observed in [6]. Classification accuracy is statistically tested using the Friedman test [25]. The null hypothesis, i.e., there is no difference in performance between the eight methods, can safely be rejected on the 0.05 significance level. However, when testing the pair wise differences with a Nemenyi test [25], no significant difference can be detected, most likely due to the limited number of datasets considered.

IV. Concluding remarks

Classification accuracy of nearest neighbor can be improved using dimensionality reduction and further improved by using different methods of feature and classifier fusion. In this paper, we have investigated two novel methods for fusing features and classifiers in conjunction with three dimensionality reduction methods for nearest neighbor classifier in high dimensions.

The new methods, i.e., FF2 and CF2, outperform raw classification accuracies in 6 and 7 datasets respectively, out of 8. If one compares the novel fusion methods with each individual dimensionality reduction method, the novel methods outperform the use of PCA and IG in a majority of the cases whereas there is a tie between using PLS and CF2. It was observed that the novel methods perform particularly well when all the dimensionality reduction methods outperform using the original feature set.

Although the fusion methods proposed in this study does not outperform the fusion methods investigated in [6] for all cases as expected, the novel classifier fusion method outperforms the previous in 4 out of 8 cases. The novel classifier fusion method also obtained the best accuracy of all methods on the NCI data set. However, the novel feature fusion method performed poorly and this might be due to that it often results in a higher number of features compared to the other methods. So even if the selected number of features is optimal when considering each dimensionality reduction method separately, the number of features obtained from combining these may be far from optimal. Hence, the idea of optimizing the number of features separately for each dimensionality reduction method can only be recommended for classifier fusion.

Acknowledgment

The first author gratefully acknowledges support through the SIDA/SAREC IT project.

References

[1] J. Quackenbush, "Microarray analysis and tumor classification," *The New England Journal of Medicine*, vol. 354, no. 23, pp. 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*, vol. 1, pp. 203–209, 2002.

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

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

[5] 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*, 2007, pp. 800–809.

[6] ——, "Fusion of dimensionality reduction methods: a case study in microarray classification," in *Proceedings of the 12th International Conference on Information Fusion*, 2009, pp. 460–465.

[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 regression (PLS-regression).* Thousand Oaks (CA): Sage, 2003, pp. 792–795.

Table II
CLASSIFICATION ACCURACIES

Data set	Raw	PCA	PLS	IG	FF1	CF1	FF2	CF2
Central Nervous	56.67	55.00	61.67	53.33	61.67	51.67	56.67	53.33
Colon Tumor	77.42	71.90	82.86	76.19	77.86	84.52	77.62	78.10
Leukemia	89.47	94.17	94.17	94.17	96.67	96.67	91.67	94.17
Prostate	85.29	83.36	88.27	90.45	92.27	93.27	92.36	91.27
Brain	76.19	76.50	90.00	72.00	73.50	76.50	78.50	81.50
Lymphoma	98.39	100.00	100.00	97.14	100.00	100.00	100.00	100.00
NCI60	68.85	68.57	72.24	62.14	73.57	68.57	68.57	75.24
SRBCT	87.30	92.14	96.90	100.00	100.00	98.33	100.00	98.57

[9] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993.

[10] StatSoft Inc., "Electronic statistics textbook," 2006, http://www.statsoft.com/textbook/stathome.html. [Online]. Available: http://www.statsoft.com/textbook/stpls.html

[11] A. Boulesteix, "PLS dimension reduction for classification with microarray data," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, 2004.

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

[13] U. M. Fayyad and K. B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning*, vol. 8, pp. 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, pp. 121–126.

[15] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," *Computing and Information Systems*, vol. 7, pp. 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*, vol. 415, pp. 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*, vol. 96, 1999, pp. 6745– 6750.

[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*, vol. 286, pp. 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*, vol. 403, pp. 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*, vol. 24, no. 3, pp. 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*, vol. 7, no. 3, 2006.

[23] W. Melssen, R. Wehrens, and L. Buydens, "Supervised kohonen networks for classification problems," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, pp. 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*, vol. 86, no. 1, pp. 102–120, 2006.

[25] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.