# Combining Feature Subsets in Feature Selection

Marina Skurichina and Robert P.W. Duin

Information and Communication Theory Group, Faculty of Electrical Engineering,
Mathematics and Computer Science, Delft University of Technology,
P.O. Box 5031, 2600GA Delft, The Netherlands
m.skurichina@ewi.tudelft.nl
r.p.w.duin@ewi.tudelft.nl

**Abstract.** In feature selection, a part of the features is chosen as a new feature subset, while the rest of the features is ignored. The neglected features still, however, may contain useful information for discriminating the data classes. To make use of this information, the combined classifier approach can be used. In our paper we study the efficiency of combining applied on top of feature selection/extraction. As well, we analyze conditions when combining classifiers on multiple feature subsets is more beneficial than exploiting a single selected feature set.

## 1 Introduction

In many medical applications it is very difficult to collect a large number of observations (for instance, patients with a certain disease). Usually the number of measurements is limited. On the other hand, such measurements can have a large number of attributes (features). By this, we face the small sample size problem: the number of measurements is smaller than or comparable with the data dimensionality. In such conditions, it is difficult (or even impossible) to construct a good classification rule [1]. One has to reduce the data dimensionality. This can be done by applying feature selection or extraction procedures to the dataset.

When using feature extraction techniques (for instance, PCA [2]), all features contribute in new extracted features. So, one may expect that all (or the most) information useful for discrimination between data classes is taken into account and reflected in an extracted feature set. However, when the feature selection approach (like forward or backward feature selection, for instance) is used, only a subset of features is chosen as a new feature set. While the rest of features (that may still be discriminative) is ignored. Useful information hidden in these features is not taken into consideration. This may result in a poor performance on the selected feature subset.

To benefit from the information presented in the neglected features in feature selection, we suggest to use the combining approach. Instead of constructing a single classifier on one selected feature set, we propose to use the combined decision of classifiers constructed on sequentially selected sets of features. First, an optimal feature set (subspace) is selected. Then on the rest of features, we select the second best feature

N.C. Oza et al. (Eds.): MCS 2005, LNCS 3541, pp. 165–175, 2005.

set etc., until all features are included in a particular feature set. By this, we have a number of optimal feature subsets and may construct a classifier on each of them. By combining decisions of these classifiers, we use all information represented in the original feature space that may improve the performance achieved on a single subset of features.

In order to demonstrate the advantage of combining multiple feature subsets and to study conditions when combining applied on top of feature selection is beneficial, we have selected four real datasets: autofluorescence spectra measured in the oral cavity, images of handwritten digits, sonar and ionosphere datasets. All datasets introduce a 2-class problem. The data are described in section 2. The description of our combined approach applied to feature selection/extraction and the results of our simulation study are presented in section 3. Conclusions can be found in section 4.

## 2 Data

We perform our study on the following four examples.

The first dataset represents autofluorescence spectra measured in the oral cavity. The data consist of the autofluorescence spectra acquired from healthy and diseased mucosa in the oral cavity. The measurements were performed at the Department of Oral and Maxillofacial Surgery of the University Hospital of Groningen [3]. The measurements were taken at 11 different anatomical locations with excitation wavelength equal to 365 nm. Autofluorescence spectra were collected from 70 volunteers with no clinically observable lesions of the oral mucosa and 155 patients having lesions in the oral cavity. Some patients suffered from multiple lesions, so that a total of 172 unique lesions could be measured. However, a number of measurement sessions had to be left out of the analysis for different reasons: 1) because an accurate diagnosis was not available for the lesion at the time of measurement, 2) because it was not insured that the probe had been located at the correct position because the lesion was hardly visible or very small, 3) because the patient had already been receiving therapy, or 4) because the diagnosis for some benign lesions was overly clear. In total, 581 spectra representing healthy tissue and 123 spectra representing diseased tissue (of which 95 were benign, 11 dysplastic and 17 cancerous) were obtained. After preprocessing [3], each spectrum consists of 199 bins (pixels/wavelengths).

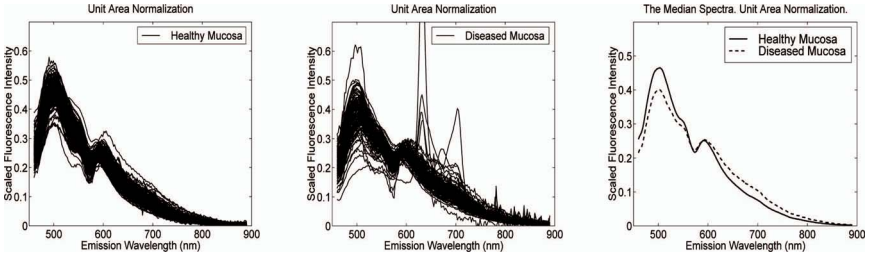


**Fig. 1.** Normalized autofluorescence spectra for healthy and diseased mucosa in oral cavity

In order to get rid of a large deviation in the spectral intensity within each data class, we normalized spectra by the Unit Area (UA)

$$a_i^{UA} = \frac{a_i}{U}, \quad U = \sum_{j=1}^{199} a_j, \quad i = 1, \ldots, 199, \tag{1}$$

where $a_i$ is an intensity of a spectrum $\boldsymbol{A}=\{a_1, ...., a_{199}\}$ at bin $i$, $i$=1,...,199. Normalized autofluorescence spectra representing healthy and diseased tissues and their median spectra are illustrated in Fig. 1.

The second dataset is handwritten digit "mfeat" dataset from [4]. Originally the data contain 10 digit classes with 200 samples per class and six different feature sets. For our study we have chosen the feature set of pixel averages consisting of 240 attributes (features). As well, we restricted ourselves to two-class problem selecting classes which represent digits 3 and 8. The example of these data classes are presented in Fig. 2.

The third dataset is "sonar" dataset from UCI Repository [4]. The task of the sonar dataset is to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock. Thus the dataset consists of two data classes. The first data class contains 111 objects obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. The second class contains 97 objects obtained from rocks under similar conditions. Each object is a set of 60 numbers in the range 0.0 to 1.0. Thus, the data are 60-dimensional. Each number (feature) represents the energy within a particular frequency band, integrated over a certain period of time.

The last dataset is also taken from the UCI Repository [4]. It is the "ionosphere" dataset. These radar data were collected by a system consisted of 16 high-frequency antennas with a total transmitted power of about 6.4kW in Goose Bay, Labrador. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of structure in the ionosphere. "Bad" returns are those that do not return anything: Their signals pass through the ionosphere. The data are described by 34 features that introduce two attributes for 17 pulse numbers corresponding to the complex values returned by the function resulting from the complex electromagnetic signal. This dataset consists of 351 objects in total, belonging to two data classes: 225 objects belong to "good" class and 126 objects belong to "bad" class.

**Fig. 2.** The example of handwritten digits "3" and "8"

For our simulation study, training datasets with 20 (for handwritten digits dataset), 50 (for spectral dataset), 15 (for sonar dataset) and 10 (for ionosphere dataset) samples per class are chosen randomly from the total set. The remaining data are used for testing. The prior class probabilities are set to be equal. To evaluate the classification performance when the combined approach applied to feature selection and standard feature selection/extraction methods are used, we have chosen for Linear Discriminant Analysis (LDA) [2] which was the best performing classifier for these applications at the given sample size. In particular, we apply the linear classifier which constructs a linear discriminant function assuming normal class distributions and using a joint class covariance matrix for both data classes. All experiments are repeated 50 times on independent training sample sets for forward feature selection, random feature selection and PCA. Additionally, for the random feature selection we repeat the random permutation and split of the feature set into the subsets 10 times. In all figures the averaged results over 50 trials (500 trials for the random feature selection) are presented and we do not mention that anymore. The standard deviation of the reported mean generalization errors (the mean per two data classes) is approximately 0.01 for each considered case.

## 3 Combining Feature Subsets in Forward and Random Feature Selection and in PCA

When the number of available observations is limited and smaller than the data dimensionality, one is forced to apply feature selection/extraction techniques in order to construct a reliable classifier to solve the problem. One of the main differences between feature selection and feature extraction approaches is in the amount of useful information they are capable to retrieve from the data representation in the feature space. In general, feature extraction techniques make use of all original data features when creating new features. The new extracted features are a combination of the original ones. By this, the new extracted feature set may contain all (or almost all, we believe) useful information for classifying the data stored in a multidimensional data representation. However, feature extraction is an operation in the high dimensional space and for small sample sizes (which may be the reason to perform feature reduction) it may suffer from overtraining. As well, it may happen that feature extraction fails due to very complex class distributions in the high dimensional feature space. In this case, feature selection may be an alternative.

Feature selection is a special case of feature extraction. In feature selection, only a part of the original features is chosen as a new feature subset. The rest of features is ignored. Sometimes (depending on data representation in the feature space), this approach works good when a few features provide a good separation between data classes and the rest of features introduces noise. However, in the case when all data features are informative without a clear preference to each other, feature selection approach may be harmful. Useful information stored in neglected features is not taken into account. The selected feature subset is not optimal. That may cause a poor performance of the classifier on this feature subset. As well, some feature selection

procedures (for instance, the forward or backward feature selection) have another drawback: the selection of features is performed sequentially one by one. The union of the first best feature selected with the second best one does not necessarily represent the best discriminative pair of features. By this, the selected feature subset might be not the most advantageous one. In addition, the efficiency of feature selection may suffer from the curse of dimensionality. When selecting features, the performance is judged by the class separability provided by these features. The separability of data classes is evaluated by some criterion. Such a criterion can be the performance of a particular classifier. However, the performance of almost all classifiers depends on the relation between the training sample size and the data dimensionality. For a finite size of the training set and an increasing feature space dimensionality, one observes that first the generalization error of a classifier decreases to its minimum and then starts to increase (see Fig. 3). The latter increase of the classification error is caused by the growing complexity of the classification rule that cannot be properly trained due to a lack of training data. In feature selection, a feature subset corresponding to a minimum classification error is chosen as the best one. It is indeed optimal by the feature size related to the available training sample size. But it is not necessary the optimal one in general. The rest of features can be still informative to discriminate the data classes. But they are not taken into consideration due to a shortage of data to construct a reliable classifier in the feature subspace of a higher dimensionality.

To overcome the drawbacks of a standard feature selection technique (say, the forward feature selection) and to make use of all information present in the original feature space when performing feature selection, we suggest to apply the classifiers combining approach on top of feature selection. Previously it has been demonstrated that combining performs well when it is applied to the data described by the different types of representations [5] or when the random subspaces of the data feature set are used [6]. Therefore, we expect that combining classifiers in selected feature spaces (on the selected subsets of features) will be also beneficial.

In our study we consider two examples of feature selection: the forward feature selection and the random feature selection. We choose the forward feature selection for our study because it is a standard well-known and relatively fast approach. On the other hand, forward selection has a number of drawbacks mentioned above. Due to them, the

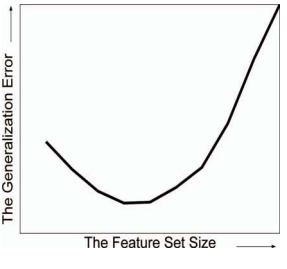


**Fig. 3.** The behaviour of generalization error for finite training sample size and increasing data dimensionality

selected feature subsets may be far from optimal. So, the random feature subsets may be as good as the specially selected subsets and they (the random subsets) do not require extensive calculations to be obtained. By this reason, we decided to consider random feature selection as well and compare its performance with the forward technique when single and sequential multiple trials of feature selection are performed. The performance of both considered feature selection techniques is also compared with the performance of the most popular feature extraction technique - Principal Component Analysis.

In the forward feature selection, we sequentially select a number of optimal feature subsets having the same size $s$. First, we perform feature selection on the entire feature set (which consists of $p$ features) obtaining the first feature subset. Then, on the rest of features $(p-s)$ (excluding already selected features) we find the next optimal feature subset. We again omit the selected features from consideration and apply the forward feature selection to the remaining $(p-2s)$ features getting the third optimal feature subset and so on, until all features are assigned to one of the selected feature subsets. All obtained feature subsets have the same dimensionality $s$ with an exception of the last one, which consists of the remaining $p-t\times s$ features after $t=\lfloor p/s \rfloor$ previously performed feature selection trials. On each of the $t+1$ selected feature subsets, the linear classifier is constructed. The decisions of these are aggregated by three different combining rules: the weighted majority voting [7], the mean rule and the decision templates (DT) [8].

For random feature selection we first randomly permute features in the original feature set and then we split the feature set into $t+1$ subsets (so, each feature is included only in one feature subset and does not appear in other subsets). By this, all obtained random feature subspaces (subsets) have the same dimensionality $s$ besides the last one with dimensionality equal to $p-t\times s$. On each of the selected random subspaces, we construct a linear classifier. Then $t+1$ obtained classifiers are combined by the weighted majority rule, the mean rule and the DT combining rule. Let us note that the random feature selection performed by us is different from the Random Subspace Method (RSM) [6]. In both, in random feature selection and in the RSM, features are selected randomly. However, the drawback of the RSM is that it is not guarantied that all features (and therefore all useful information) are taken into consideration at the end (each time one selects a random feature subset from the total set of features but not from the rest after previous selections). Hence, some features may be multiply represented in feature subsets and some may not be taken into consideration at all (especially when a limited number of small random subspaces is used). In our study it is important that all features are used once when we apply combining on top of feature selection. The RSM does not satisfy this requirement. For this reason, we do not include this technique in our study.

When applying the combining technique on top of PCA, we split the set of principal components into subspaces as following. The first feature subset consists of the first $s$ principal components, the second one contains the next $s$ principal components and so on. Similar to the previous cases, the weighted majority, the mean rule and the DT combining rule are used to aggregate classifiers constructed on the subsets of principal components.

The performance of the linear classifier on a single selected feature subset and the combined decision of linear classifiers on multiply selected feature subsets for forward and random feature selection are illustrated in Fig. 4 on the examples of the spectral and digit datasets and in Fig. 5 for the sonar and ionosphere datasets. We see that the benefit of the combining approach applied on top of feature selection depends on the data distribution, on the type of feature selection and on the size of feature subsets used. In the majority of cases, the better performance is achieved by combining feature subsets than by exploiting a single selected/extracted feature subset. Combining is more effective when it is applied on top of a "weak" feature selection technique (random and forward feature selection) than on top of a "strong" feature selection/extraction technique (in our case PCA). The most improvement in performance is gained on the feature subset sizes that are approximately more than twice smaller than the training sample size.

However, no large difference is noticed between random and forward feature selection when combining is applied to multiple feature subsets. It might be explained by the fact that in both cases all original features participate in the combined decision. The random selection approach seems to be more attractive than the forward feature selection technique by two reasons. First, it might be more successful in selecting independent feature subsets than the forward feature selection, for instance for datasets with many correlated features like spectral data. Then, we may obtain independent classifiers on the feature subsets, which for combining are more beneficial than combining correlated classifiers [9]. Secondly, the random feature selection is very fast and does not need any sophisticated algorithm for finding an optimal feature subset.

In our examples, the feature extraction approach (PCA) performs better (with exception of very small sizes of exploited feature subsets) than a single trial of forward or random feature selection (see Fig. 4 and 5), because the extracted features contain more information for discrimination between data classes than a single selected feature set. What concerns the combining approach applied on top of PCA, its' success merely depends on the data distribution and on the size of feature subsets used. Exercising the classifiers combining on principal components may be advantageous only for small feature subset sizes (that are approximately twice smaller than the training sample size) when the subset of the first few extracted principal components is too small and does not preserve enough useful information to discriminate between data classes. For some datasets (for instance, for the spectral and digit data, see Fig. 4) PCA succeeds in extracting good features. In these cases, the combining approach is useless: a single classifier constructed on a sufficient number of the first principal components performs better than combining of sequential subsets of principal components. However, for other datasets (e.g., the sonar and ionosphere data, see Fig. 5) combining applied on top of PCA and performed on small feature subsets is very effective: it improves the best performance achieved by PCA using a single feature subset. Interestingly, for datasets like sonar and ionosphere (see Fig. 5), using combining on top of random feature selection is even more beneficial than PCA (with or without applying the combining approach to subsets of principal components) when small feature subspaces are considered.

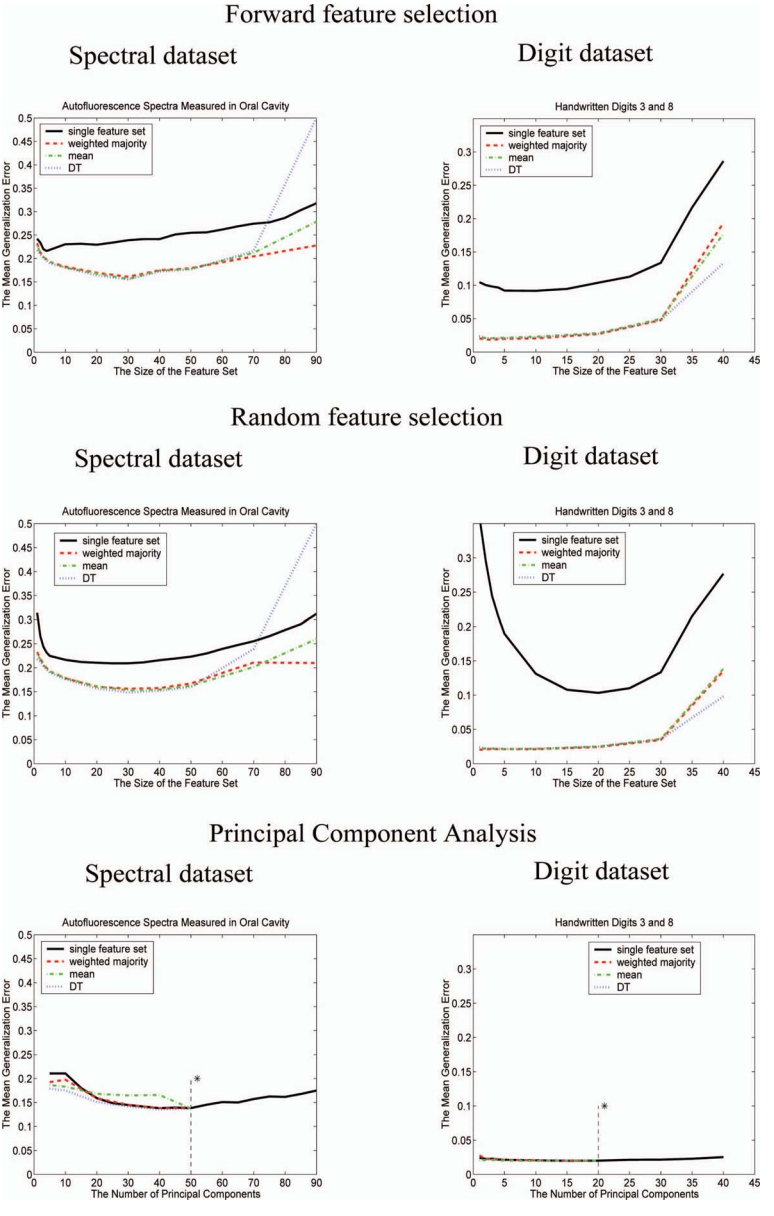


**Fig. 4.** The mean generalization error (GE) of a single and combined LDA for the forward feature selection (top plots), the random feature selection (middle plots) and PCA (bottom plots) for the spectral (50+50 training objects) and digit (20+20 training objects) datasets. *)For PCA, 100 and 40 principal components are possible to retrieve, because training sample size equals to 100 and 40 objects for spectral and digit dataset, respectively. Hence, the classifiers combining on principal components is performed only up to the feature set size is equal to 50 for spectral dataset and to 20 for digit dataset

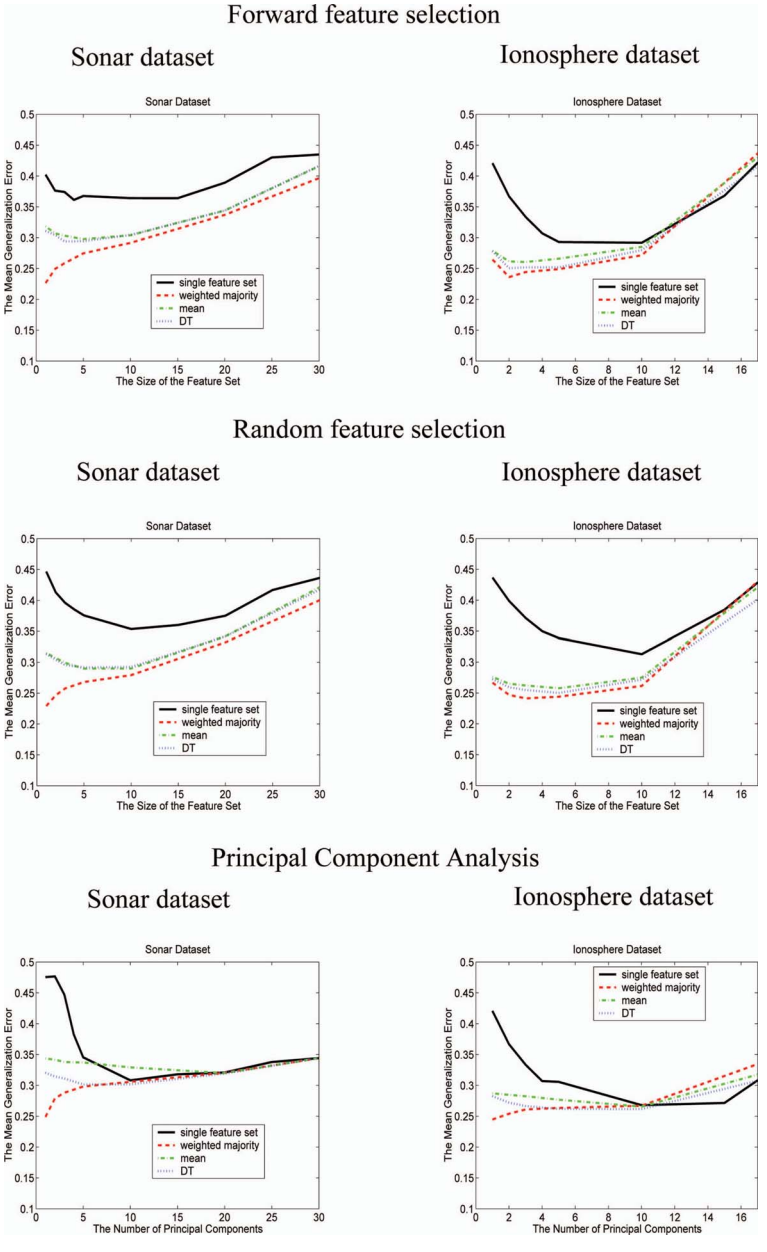


**Fig. 5.** The mean generalization error (GE) of a single and combined LDA for the forward feature selection (top plots), the random feature selection (middle plots) and PCA (bottom plots) for the sonar (15+15 training objects) and ionosphere (10+10 training objects) datasets. For PCA, 30 and 20 principal components are possible to retrieve, because training sample size equals to 30 and 20 objects for sonar and ionosphere dataset, respectively. Hence, the classifiers combining is performed only up to the feature set size is equal to 30 for sonar dataset and to 20 for ionosphere dataset

## 4 Conclusions

In order to construct reliable classifiers for high dimensional datasets with a limited number of observations, it is needed to reduce the data dimensionality. Feature selection or feature extraction can be considered. In feature selection only a part of the features is taken into consideration, while the remaining features (that may be still informative) are neglected. To benefit from this information we have suggested to apply the classifier combining approach on top of feature selection/extraction. We have found that the success of combining feature subsets depends on the data distribution, on the type of feature selection/extraction and on the size of feature subsets used.

The combining approach applied on top of feature selection/extraction is the most effective when using small feature subsets. Combining feature subspaces is more beneficial for weak feature selection techniques (like forward or random feature selection) than for strong feature extraction techniques (like PCA).

We have found that exercising the classifiers combining on the subsets of features results in a similar performance for both forward and random feature selection techniques. Forward feature selection does not seem to be the optimal approach to select the best possible feature subsets especially for datasets with small sample sizes. By this, when combining multiple feature subsets, random feature selection may be preferred to the forward feature selection as it is fast and might be more successful in obtaining independent feature subsets (that may result in a more beneficial ensemble of independent classifiers).

When the feature extraction approach is used, all original features contribute in a new extracted feature set. By this, feature extraction technique like PCA is more advantageous than weak feature selection techniques (like forward or random selection). Depending on the data distribution, one may need quite many principal components in order to obtain a good performing classification rule. In the case of small sample sizes, it is not always possible. In such a case, it might be useful to apply combining on top of feature extraction. However, the combining approach on top of PCA is not always useful. It is beneficial only when small feature subsets are exploited and for datasets where the first principal components fail in good discrimination between data classes.

## Acknowledgment

This work was supported by the Dutch Technology Foundation (STW), grant RRN 5316.

## References

1. Jain, A.K., Chandrasekaran, B.: Dimensionality and Sample Size Considerations in Pattern Recognition Practice. In: Krishnaiah, P.R., Kanal, L.N. (eds.): Handbook of Statistics, Vol. 2. North-Holland, Amsterdam (1987) 835-855
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (1990) 400-407

3. De Veld, D.C.G., Skurichina, M., Witjes, M.J.H., et.al. Autofluorescence and Diffuse Reflectance Spectroscopy for Oral Oncology. Accepted in Lasers in Surgery and Medicine (2005)
4. Blake, C.L., and Merz, C.J.: UCI repository of machine learning databases (1998). http://www.ics.uci.edu/~mlearn/MLRepository.html
5. Tax, D.M.J., van Breukelen, M., Duin, R.P.W. and Kittler, J.: Combining Multiple Classifiers by Averaging or Multiplying? Pattern Recognition, Vol. 33(**9**) (2000) 1475-1485
6. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20(**8**) (1998) 832-844
7. Freund, Y., and Shapire, R.E.: Experiments with a New Boosting Algorithm. Proceedings of the 13th International Conference on Machine Learning (1996) 148-156
8. Kuncheva, L.I., Bezdek, J.C., and Duin, R.P.W.: Decision Templates for Multiple Classifier Fusion: An Experimental Comparison. Pattern Recognition, Vol. 34(**2**) (2001) 299-314
9. Kuncheva, L.I.: Combining Pattern Classifiers. Methods and Algorithms. Wiley (2004)