

A Robust Subspace Classifier*

Horst Bischof[†], Aleš Leonardis[‡] and Florian Pezzeri[†]

[†] Pattern Recog. and Image Proc. Group
Vienna University of Technology
A-1040 Vienna, Austria
{bis,fpezz}@prip.tuwien.ac.at

[‡] Faculty of Computer and Info. Science
University of Ljubljana
SI-1001 Ljubljana, Slovenia
alesl@fri.uni-lj.si

Abstract

In this paper we study the problem of missing features and the issues of robustness of subspace classification methods. We propose a new robust method for subspace classification which can cope with missing features and/or outliers. The main idea of our method is to use a robust projection of the patterns onto a subspace. We demonstrate our approach on Cervicomotography data and compare our results to the results obtained by using various decision tree algorithms.

1. Introduction

The amount of data that is collected and archived everyday has increased dramatically over the last few years. This has also increased the need for automatic processing of data in order to turn it into a valuable information. Pattern recognition techniques provide the necessary tools for that purpose. However, with the increasing amounts of data, the demands on pattern recognition systems also keep raising. One of the major problems that we have to deal with is that the dimensionality of the data is quite high cf. (*curse of dimensionality*) while at the same time there is a limited number of training samples available. Secondly, due to the automatic acquisition of the data, some features might be missing and others might be wrong (e.g., outliers). A successful pattern recognition method has to handle these problems.

In this paper we propose how to solve the problem of missing features and how to increase the robustness of subspace classification methods. The subspace pattern recognition method [7, 12] is a statistical method proposed originally by Watanabe. In subspace classification, each pattern class is represented by a subspace spanned by a group of

basis vectors — the orthogonal components obtained by the principal component analysis. This method corresponds to the least squares fit on the given data. To account for the non-gaussian noise, the basic subspace method has been modified to the learning subspace method (LSM) [4] and the averaged learning subspace method (ALSM) [7]. Recently Parkash & Murthy have proposed the Hebbian learning subspace method [10] and the growing subspace methods [9]. The extended adaptive learning subspace method (EALSM) [9] uses multiple PCA-subspaces per class.

However, all these methods cannot handle the problems due to missing data (features) and/or outliers in the classification phase¹. On the other hand, several techniques have been proposed to handle the missing feature problem in other classification methods, e.g., [6]. A common practice is to replace the missing values by the mean value of the corresponding variable. However, this approach has serious shortcomings (see e.g., [1]). A more elaborate approach is to express any variable which has missing values in terms of a regression over the other variables, and then to use the regression function to fill the missing values. This approach tends to cause problems as it underestimates the covariance in the data since the regression function is assumed to be noise-free. Another possibility is to use the EM algorithm [2] to estimate both the missing values and the probability density of the data. However, for high dimensional feature spaces it is infeasible to estimate the densities. The problems with missing data get further complicated when our data set also contains a number of wrong measurements (i.e., outliers). In this case we first have to detect the outliers and then to treat those features as missing values.

In this paper we introduce a robust method for subspace classification which can cope with missing features and/or outliers. The main idea of our method is to use a robust projection to calculate the coefficients of the principal components.

*This work was supported by a grant from the Austrian National Fonds zur Förderung der wissenschaftlichen Forschung (No. S7002MAT). We are grateful to Dr. Berger and Dr. Lechner-Steinleitner from the Univ. of Innsbruck for providing the CMG-data set. A. L. acknowledges the support from the Ministry of Science and Technology of Republic of Slovenia (Project J2-8829) and by U.S.-Slovene Joint Board (Project #95-158).

¹for the training phase robust principle component analysis might be used, e.g., [13].

2. Subspace Classification

The subspace classification methods are decision theoretic pattern recognition methods, where the primary model for a class is a linear subspace of the Euclidean pattern space². Any set of m linearly independent vectors $\mathbf{p}_1, \dots, \mathbf{p}_m \in \mathbb{R}^n$, ($m < n$) spans a subspace L : $L = L(\mathbf{p}_1, \dots, \mathbf{p}_m) = \{\mathbf{x} | \mathbf{x} = \sum_{i=1}^m y_i \mathbf{p}_i, y_i \in \mathbb{R}\}$. In what follows we assume that the vectors $\mathbf{p}_i$ are obtained by principal component analysis and are therefore orthonormal, i.e., $\mathbf{p}_i^T \mathbf{p}_j = \delta_{ij}$. We can project a vector $\mathbf{x}$ on the subspace L by: $\mathbf{y} = \mathbf{x}^T \mathbf{P}$, where $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_m]$ and $\mathbf{y} \in \mathbb{R}^m$. To reconstruct the vector from this projection we can calculate:

$$\hat{\mathbf{x}} = \sum_{i=1}^m y_i \mathbf{p}_i = \sum_{i=1}^m (\mathbf{x}^T \mathbf{p}_i) \mathbf{p}_i . \quad (1)$$

The orthogonal residual $\tilde{\mathbf{x}}$ of this projection is then given by: $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$. Therefore, any vector $\mathbf{x}$ can be written as, $\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}}$. The classification criterion for pattern $\mathbf{x}$, according to the subspace classifier is

if $d(\mathbf{x}, L^{(i)}) > d(\mathbf{x}, L^{(j)})$, $i \neq j$ then classify $\mathbf{x}$ in class $\omega^{(i)}$,

where $L^{(i)}$ is the subspace associated with the class $\omega^{(i)}$, and $d(\mathbf{x}, L) = \sum_{i=1}^m y_i^2 = \sum_{i=1}^m (\mathbf{x}^T \mathbf{p}_i)^2$ is the squared orthogonal projection distance.

The subspaces for the classes $L^{(i)}$ are usually determined by principle component analysis, i.e., we choose those eigenvectors of the class correlation matrix that correspond to the first m largest eigenvalues such that a prespecified percentage of the class variance is captured. In the LSM and ALSM subspace methods, the obtained eigenvectors are rotated by a learning algorithm in order to decrease the misclassification.

2.1. Robust Subspace Classification

Since $\hat{\mathbf{x}}$ is the unique vector in L which minimizes $\|\mathbf{x} - \hat{\mathbf{x}}\|^2$, $\hat{\mathbf{x}} \in L$ (see [7]), where $\|\cdot\|$ is the Euclidean norm, the classification is based on the quadratic error which is not robust; in other words, the result of classification is very sensitive to outliers [3, 11]. It is easy to show that just by modifying one component of $\mathbf{x}$ (i.e., one feature) we can get arbitrarily wrong projection results (see [5]). Similarly, we can argue that also missing features cause problems in the classification.

In order to solve these problems we propose a robust way of calculating the projection. Let us consider again Eq. (1): $\hat{\mathbf{x}} = \sum_{i=1}^m y_i \mathbf{p}_i$. To determine the coefficient vector $\mathbf{y}$ we can set up the following linear least squares problem:

²To simplify the discussion we deal only with a single subspace per class, however the extension to multiple subspaces is straight forward.

$$\mathbf{x} = \sum_{i=1}^m y_i \mathbf{p}_i = \mathbf{y}^T \mathbf{P} . \quad (2)$$

The least squares solution of Eq. (2) is given by: $\mathbf{y}_{LS}^T = \mathbf{P}^\dagger \mathbf{x}$, where $\mathbf{P}^\dagger = (\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T$ is the pseudoinverse. Since $\mathbf{P}$ is orthonormal, it follows that $\mathbf{P}^\dagger = \mathbf{P}^T$ and therefore $\mathbf{y}_{LS} = \mathbf{x}^T \mathbf{P}$ which is also the solution of the projection equation. Since we are interested in a robust solution, we solve Eq. (2) not in the least squares sense but rather in a robust manner. In principle we can use any robust approach to the solution of the equation (see [11]). In particular, we use the following method:

- We select a random subset of features from $\mathbf{x}$. Let $\mathbf{x}'$ and $\mathbf{p}'_i$ be the reduced vector and the correspondingly reduced basis vectors, respectively. First, we solve $\mathbf{x}' = \mathbf{y}^T \mathbf{P}'$ in the least squares sense³ and then we iteratively, based on the error distribution, reduce the number of features, which yields the final solution.
- In order to increase further the tolerance for the number of outliers, we perform the above procedure on multiple random subsets and select as the final solution the one with the largest $d(\mathbf{x}, L)$.

This robust procedure is effective when the number of features is much larger than the number of eigenvectors per class.

In a similar way we can deal with the problem of missing features. In this case, $\mathbf{x}'$ is constructed such that exactly the missing features are not taken into account. Again, this yields a reduced matrix $\mathbf{P}$ which is no longer orthogonal. Then we proceed as in the robust case. When we know that there are no outliers present in the data, we just solve the equation in the least squares sense without iteration.

3. Experimental Results

We have tested the robust subspace classifier on various data sets, however, due to lack of space, only the results on the Cervicomotography data are reported here. The subspaces are calculated using principal component analysis on the training set. For each class one subspace is calculated. The number of principal components is selected equally for all classes, such that at least 95% of the class variance is covered. We compare the results to various decision tree algorithms.

As a training algorithm we have used the principle component analysis despite the fact that using a learning subspace method (e.g., ALSM) would lead to increased classification accuracy. Namely, the goal of this paper is to

³Note that due to the reduction of the basis vectors, $\mathbf{P}$ is no longer orthogonal.

demonstrate that our method is robust and can deal with missing features. The results we draw on the robustness of the classifier are independent of the training method.

The cervicomotography data (CMG) was obtained from the University of Innsbruck, where a special measurement device for measuring the three dimensional head movement of patients was developed. This device is used by medical doctors to determine if a patient has whiplash, an injury very common after car accidents. The patient is asked to perform specific head movements which are registered by the CMG device. From the obtained curves the doctors can assess the injury of the patient. Since the judgment of a doctor is subjective, it is desirable to have an automatic and objective classification system. In order to classify the data automatically, 129 numeric features have been extracted from the curves (e.g., various Fourier coefficients, max/min, standard deviation of amplitude, etc., see [8] for details). The data set used consists of 764 cases (172 of healthy people and 592 of people with different degrees of whiplash). On the average 5.1% of the features of each pattern are missing in the data set. In a previous study [8] different decision tree algorithms like ID3, C4.5 and OC1 have been compared on this data set. The missing features were replaced by the mean value of the features. Fig. 1 shows the average results of 10-fold cross validation obtained with three decision tree methods. Also, feature subset selection (FSS) has been used on this data set but did not improve the results significantly, as can be seen from Fig. 1.

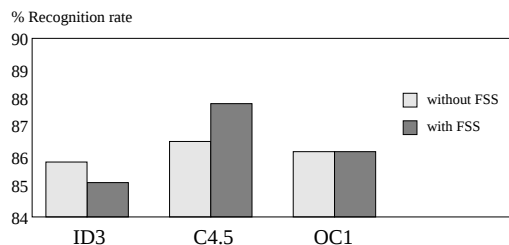


Figure 1. Results of decision tree methods on CMG data

We tested our method on this dataset using the 10 eigenvectors corresponding to the 10 largest eigenvalues of the class covariance matrix per class. For the training set we replaced the missing features by the mean values of the respective class. For classification, we eliminated from the eigenvectors the values corresponding to missing features. The projection distance was then calculated from this reduced pattern (we did not use multiple subsets per vector, nor the robust equation solver) as explained in section 2.1. The average classification accuracy is 96.73% which is approximately 9% better than the results delivered by the best decision tree method. This clearly demonstrates the superi-

ority of our method.

4. Conclusions

In this paper we have introduced a robust subspace classifier which can deal with missing features and/or outliers. We have shown that the usual projection operation of subspace classifiers can be replaced by a robust method. We have demonstrated the effectiveness of our approach on the CMG data. We have shown that our method effectively deals with the problem of missing features and outperforms decision tree algorithms.

Our emphasis in this paper was on the robustness during the classification phase and not on the robustness of classifier construction (learning). The method we have introduced is most effective when the dimensionality of the feature vector is high while at the same time the dimensionality of the subspaces is low. This means that there is a lot of redundancy in the data which is required in order to achieve a robust method.

References

- [1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [2] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J.R. Statist. Soc. B*, 39:1–38, 1977.
- [3] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.
- [4] T. Kohonen. *Self-Organization and Associative Memory*. Berlin: Springer-Verlag, 1989.
- [5] A. Leonardis and H. Bischof. Dealing with occlusions in the eigenspace approach. In *Proc. of CVPR96*, pages 453–458. IEEE Computer Society Press, 1996.
- [6] R. Little. Regression with missing X's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.
- [7] E. Oja. *Subspace Methods of Pattern Recognition*. Research Studies Press, 1983.
- [8] F. Pezzei. Vergleich von Klassifikationsalgorithmen an CMG Daten. Master's thesis, Techn. Univ. Wien, 1997.
- [9] M. Prakash and M. Murty. Growing subspace pattern recognition methods and their neural-network models. *IEEE Trans. on Neural Networks*, 8(1):161–168, 1997.
- [10] M. Prakash and M. Murty. Hebbian learning subspace method: A new approach. *Pattern Recognition*, 30(1):141–149, 1997.
- [11] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.
- [12] S. Watanabe and N. Pakvasa. Subspace method in pattern recognition. In *Proc. Int. Joint Conference on Pattern Recognition*, pages 25–32, 1973.
- [13] L. Xu and A. Yuille. Self-organizing rules for robust principle component analysis. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 467–474. Morgan Kaufmann, California, 1993.