Feature extraction for one-class classification

David M.J. Tax and Klaus-R. Müller

Fraunhofer FIRST.IDA, Kekuléstr.7, D-12489 Berlin, Germany
e-mail: davidt@first.fhg.de

Abstract. Feature reduction is often an essential part of solving a classification task. One common approach for doing this, is Principal Component Analysis. There the low variance directions in the data are removed and the high variance directions are retained. It is hoped that these high variance directions contain information about the class differences. For one-class classification or novelty detection, the classification task contains one ill-determined class, for which (almost) no information is available. In this paper we show that for one-class classification, the low-variance directions are most informative, and that in the feature reduction a bias-variance trade-off has to be considered which causes that retaining the high variance directions is often not optimal.

1 Introduction

Feature reduction is important when we want to fit a classifier using finite sample sizes. Using too many features will introduce too much noise, and classifiers can easily overfit. To avoid this, the data is preprocessed to remove as many noisy or redundant features as possible. A typical preprocessing is Principal Component Analysis (PCA), where the directions with high variance in the data are retained [Jol86]. However this heuristic does not directly guarantee that the best classification performance can be obtained.

When we are interested in the problem of novelty detection, or one-class classification [MH96,RGG95,SPST+99,JMG95,Tax01], just one class is sampled well: the target class. A classifier should be trained such that it distinguishes this class from all other possible objects, the outlier objects. To avoid the trivial solution to classify all objects as target objects, one has to assume an outlier distribution. In this paper we will assume a uniform distribution (where in practice some upper and lower bounds on the outlier domain have to be defined). In that case, the classifier should capture all (or a pre-specified fraction) of the target data, while it covers a minimum volume in the feature space.

In this paper we want discuss PCA preprocessing for this one-class classification (OCC) problem. We found that there is a bias-variance trade-off [GBD92,Hes98] in feature extraction which can cause that the low-variance directions are more useful than the high-variance directions. Removing these low-variance directions is then counter-productive. In sections 2 and 3 we look how OCC behaves on a Gaussian target distribution, and what happens when PCA is applied. In sections 4 and 5 some experiments are shown and conclusions are given.

2 OCC for Gaussian target distribution

Assume that we are given a p-dimensional dataset $\mathcal{X}^{tr} = \{\mathbf{x}_i, i = 1, .., n\}$, drawn from a Gaussian distribution $N(\mu, \Sigma)$. Assume further that the outliers are uniformly distributed in a box with the center at μ and edges of length M, where M is much larger than any of the eigenvalues of Σ. A Gaussian one-class classifier estimates the mean and covariance matrix, $\bar{\mathbf{x}}$ and S_n respectively, and uses the Mahalanobis distance $(\mathbf{x} - \mu_n)^T S_n^{-1}(\mathbf{x} - \mu_n)$ as the fit to the target class. A threshold $\theta_{p,n}$ is defined, such that the empirical target acceptance rate equals a predefined fraction, for instance $f = 90\%$. New instances $\mathbf{z}$ are evaluated by:

$$\text{accept } \mathbf{x} \quad \text{if} \quad (\mathbf{x} - \mu_n)^T S_n^{-1}(\mathbf{x} - \mu_n) \leq \theta_{p,n} \tag{1}$$

To evaluate the dependence of the performance on the dimensionality p and the sample size n, the error on the outliers or the volume captured by the classifier (false acceptance rate, f_o) and the error on the target objects (false rejection rate, f_t) have to be considered.

The captured volume f_o is determined by the ellipsoid, characterized by the (estimated) covariance matrix and the threshold $\theta_{p,n}$. From geometry we know that the volume of the ellipsoid is the volume of the sphere times the absolute value of the determinant of the transformation matrix. In total, the volume of the captured space is computed by:

$$V = \sqrt{|S_n|}\theta_{p,n}^{p/2} V_p \tag{2}$$

where V_p is the volume of a p-dimensional unit ball: $V_p = \frac{2\pi^{p/2}}{p\Gamma(p/2)}$ and $|S_n|$ the determinant of the (estimated) covariance matrix $S_n = \frac{1}{n}\sum_i (\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{x} - \bar{\mathbf{x}})$. The distribution of the determinant $|S_n|$ is the same as $|A|/(n-1)^p$, where $A = \sum_{i=1}^{n-1} \mathbf{z}_i' \mathbf{z}_i$ and $\mathbf{z}_i, i = 1, .., n-1$ are all distributed independently according to $N(0, \Sigma)$. The hth moment of $|A|$ is given by [And84]:

$$2^{hp}|\Sigma|^h \frac{\prod_{i=1}^{p} \Gamma[\frac{1}{2}(N-i)+h]}{\prod_{i=1}^{p} \Gamma[\frac{1}{2}(N-i)]} \tag{3}$$

Thus the first moment is $E(|A|) = |\Sigma| \prod_{i=1}^{p}(n+1-i)$.

The threshold $\theta_{p,n}$ can be derived from the (Mahalanobis) distance distribution of objects $\mathbf{x}_i$ to the estimated mean $\bar{\mathbf{x}}$, assuming that the objects are drawn from a Gaussian distribution. The (scaled) distances are distributed as a beta distribution [Wil62,HR99]: $\frac{(n-1)^2}{n} d_S^2(\mathbf{x}_i, \bar{\mathbf{x}}) \sim Beta\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right)$. The threshold $\theta_{p,n}$ is set such that a certain fraction f (say $f = 0.9$) of the data is accepted. This means:

$$\frac{(n-1)^2}{n} \int_0^{\theta_{p,n}} Beta\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right) = f \tag{4}$$

When we are interested in the *fraction* of the outliers which is accepted (f_o, the fraction false positive), this volume has to be compared with the volume

covered by the outlier objects M^p:

$$f_o = \frac{V}{M^p} = \sqrt{|S_n|}V_p \left(\frac{\sqrt{\theta_{p,n}}}{M} \right)^p = V_p \prod_{i=1}^{p} \frac{\sqrt{\hat{\lambda}_i \theta_{p,n}}}{M} \tag{5}$$

where $\hat{\lambda}_i$ are the eigenvalues of S_n.

Note that by the assumption that the outliers cover the whole target distribution, $\sqrt{\theta_{p,n}} < M$ and thus $\left(\frac{\sqrt{\theta_{p,n}}}{M} \right)^p$ vanishes for increasing dimensionality p. Thus the volume of the captured volume (the area for which $\mathbf{x}' S^{-1} \mathbf{x} \leq \theta_{p,n}$) and the error on the outliers will decrease.

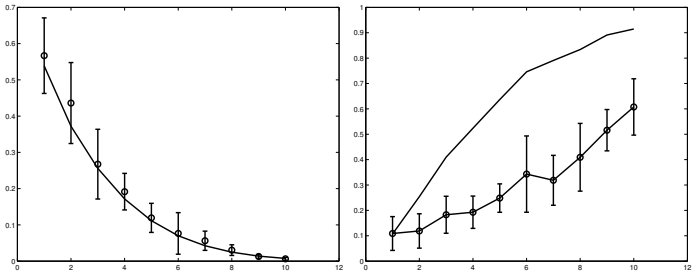

Fig. 1. Left, the relative space captured by the Gaussian one-class classifier for varying dimensionalities, $n = 25$. Right, the probability mass missed by the classifier.

In the left subplot of figure 1 the relative volumes captured by the Gaussian one-class classifier is plotted for varying p. The target class is distributed as a standard Gaussian distribution with $n = 25$ objects. The outlier objects are uniformly distributed in a box with edges of length 8. The solid line shows the theoretical fraction of accepted outlier objects (by equation (5)). The circles show the results obtained by simulation.

To estimate the error on the target set f_t, we have to compute the probability mass captured by the estimated classifier:

$$1 - f_t = \int\limits_{(\mathbf{x}-\bar{\mathbf{x}})' S_n^{-1} (\mathbf{x}-\bar{\mathbf{x}}) \leq \theta} \frac{1}{Z} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) d\mathbf{x} \tag{6}$$

To simplify this integration, we assume that that the means are estimated well and that they are at the origin, i.e. $\mu = \bar{\mathbf{x}} = \mathbf{0}$.[1] By transforming $\mathbf{y} = S_n^{-1/2} \mathbf{x}$

[1] In practice this does not completely hold, but it appears that the errors introduced by poor estimates of $|S_n|$ are much more important.

the integration area becomes circular, and we can write:

$$1 - f_t = \int\limits_{\mathbf{y}'\mathbf{y} \leq \theta} \frac{1}{Z'} \exp\left(-\frac{1}{2}\mathbf{y}^T (S_n^{T/2} \Sigma^{-1} S_n^{1/2})\mathbf{y}\right) d\mathbf{y} \tag{7}$$

where the Jacobian $|\det(S_n)|$ is absorbed in Z'.

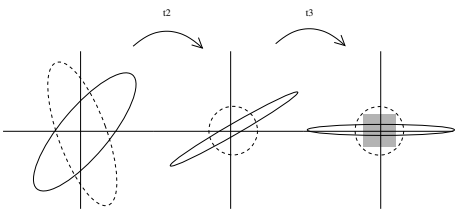

Fig. 2. Transformation of the integral (6).

Finally, we can rotate such that the main axes of $\tilde{\Sigma} = S^{T/2}\Sigma^{-1}S^{1/2}$ are aligned with the coordinate axes, by observing that we can factorize $\tilde{\Sigma} = U^T D U$ where U is an orthogonal matrix (containing the eigenvectors of $\tilde{\Sigma}$ and D is a diagonal matrix (containing the eigenvalues $\lambda_1, ..., \lambda_p$), see also figure 2 for a visualization:

$$1 - f_t = \int\limits_{\mathbf{z}'\mathbf{z} \leq \theta_{p,n}} \frac{1}{Z''} \exp\left(-\frac{1}{2}\mathbf{z}^T D \mathbf{z}\right) d\mathbf{z}$$

Unfortunately, this integral is also not analytically solvable (due to the fact that $\exp\left(-\frac{1}{2}\mathbf{z}^T D\mathbf{z}\right)$ is not rotational symmetric). We approximate this integral by integrating over a box which is within the sphere. The edges of this box have size $2a = \sqrt{\theta_{p,n}/p}$.[2] Integrating over the box, yields:

$$1 - f_t = \int_{-\sqrt{\theta_{p,n}/p}}^{\sqrt{\theta_{p,n}/p}} \frac{1}{Z''} \exp\left(-\frac{1}{2}\mathbf{z}^T D \mathbf{z}\right) d\mathbf{z} \tag{8}$$

$$= \prod_{i=1}^{p} \left(\mathrm{erf}(\sqrt{\theta_{p,n}/(p\lambda_i)}) - \mathrm{erf}(-\sqrt{\theta_{p,n}/(p\lambda_i)})\right) \tag{9}$$

In the right subplot of figure 1 f_t and the simulation results are shown for varying dimensions. It is clear that equation (9) is a crude approximation to (6), but it shows a similar trend. The target error depends on the threshold $\theta_{p,n}$ and the eigenvalues of $S_n^{T/2}\Sigma^{-1}S_n^{1/2}$, $\lambda_1, ..., \lambda_p$. When one of the eigenvalues λ_i is large, one element in the product of (9) becomes small and the $1 - f_t$ becomes small.

The total error is thus a combination of f_o, (5) and f_t, (9). Note that for the computation of (9) both the estimated as the true covariance matrix of the target class is required. This means that these equations cannot be used to find the generalization error for a given dataset.

3 Feature extraction

When the dimensionality is reduced using PCA, both high or low variance directions can be removed, effectively removing high or low eigenvalues λ_i from

[2] Integrating over the box around the sphere is also possible, but this approximation is much worse.

S_n. Equation (2) shows that $(\sqrt{\theta_{p,n}}/M)^p$ is an important factor for f_o (directly depending on p). But the actual value does not change when the difference between using the high- or low-variance directions is considered. For that, $|S_n|$ (using equation (3)) should be considered. When low variance directions are removed from $|S_n|$, $|S_n|$ basically increases, and thus f_o increases. On the other hand, when high variance directions are removed, $|S_n|$ is still mainly determined by the smallest eigenvalues, and thus the error stays constant.

When we assume that S_n approximates Σ, equation (9) shows that removing high-variance directions from $|S_n|$ will remove terms with large λ_i from $|S_n\Sigma|$ and thus small value of $\text{erf}(\sqrt{\theta/(p\lambda_i)}) - \text{erf}(-\sqrt{\theta/(p\lambda_i)})$. Effectively, $1 - f_t$ will increase and thus the error decrease. When low-variance directions are removed, the change of the product is very small. When S_n does not approximate Σ very well, removing a high- or low-variance direction gives a random result for $|S_n\Sigma|$, and so the improvement for removing high-variance directions will disappear.

In figure 3 a schematic picture for a two dimensional case is shown. The target data is distributed in the ellipsoidal area. When the data is mapped onto the largest eigenvector (direction λ_1), essentially all data between the dashed lines will be labeled target object. All outlier objects in A_1 will therefore be misclassified. When the data is mapped onto the smallest eigenvector (λ_2), all data between the solid lines (A_2) will be labeled target. Because the volume of A_2 is much smaller than that of A_1, f_o will be smaller in the second case. The f_t will be the same in both cases, so the total error decreases.

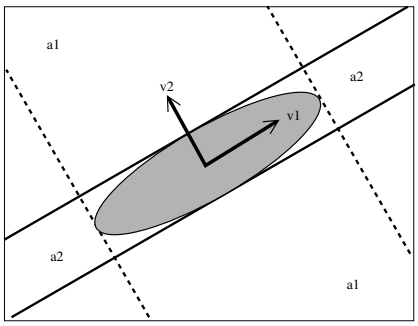

Fig. 3. Feature reduction for a two dimensional dataset.

For low sample sizes, the estimates of the low and high variance directions become noisy. In that case, not only the volume difference between A_1 and A_2 becomes smaller, also the error on the target set will in both cases increase. In figure 4 the total error is shown when the dimensionality is reduced. It shows that when data is reduced to low dimensionalities (1D, 2D), it is always better to remove the high variance directions. To determine the optimal dimensionality and which directions to use, the basic bias-variance dilemma has to be considered: how well can the different variance directions be estimated from a finite sample, and how well do these directions distinguish be-

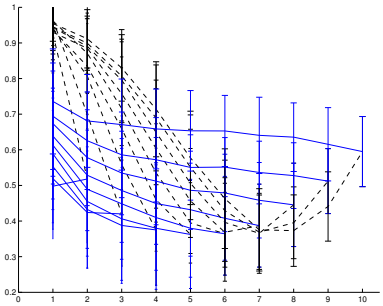

Fig. 4. The total error $f_t + f_o$.

tween the targets and outliers? We just saw that for one-class classification, the low-variance directions in the data have lowest average error (low bias). When the sample size is sufficiently high and the eigenvectors of S_n can be estimated well, the best approach is thus to reduce the dimensionality by removing high variance directions. In these cases the effect of removing a large eigenvalue from $S_n\Sigma^{-1}$ dominates the effect of removing a large eigenvalue of S_n. On the other hand, when the matrix S_n is a poor approximation to Σ, then $S_n\Sigma^{-1}$ is far from identity. Removing small eigenvalues from S_n does not mean removing small eigenvalues from $S_n\Sigma$ and the removal of a low-variance direction will still result in a large increase of f_o in equation (9).

Unfortunately, the (approximation to the) target error estimation still depends on both the estimated and 'true' covariance matrix, S_n and Σ. To estimate the dependence to the dimensionality and the sample size, datasets with known Σ and simulated S_n are generated. We use a Gaussian distributed target class, where Σ has eigenvalues between 2 and $\frac{1}{2}$, and $|\Sigma| = 1$. From each generated dataset S_n is estimated, and both f_t and f_o are computed.

Table 1. Optimal dimensionality p^* for varying sample sizes n and starting dimensionalities ($p = 2 - 10$), for both removing the high variance directions (high) and low-variance directions (low). Underlined dimensionalities give superior performance.

Variance between 1.5 and 0.66

n	10		25		50		100	
p	high	low	high	low	high	low	high	low
2	2	2	1	2	1	2	1	2
3	3	3	3	3	2	3	2	3
4	4	3	4	4	4	4	3	4
5	5	4	5	5	5	5	5	5
6	6	4	6	6	6	6	6	6
7	7	5	7	6	7	7	7	7
8	8	5	8	7	8	8	8	8
9	8	5	9	7	9	8	9	9
10			10	8	10	9	10	10

Variance between 2 and 0.5

n	10		25		50		100	
p	high	low	high	low	high	low	high	low
2	1	2	1	2	1	2	1	2
3	2	3	2	3	2	3	2	3
4	2	4	2	4	3	4	3	4
5	2	4	2	5	4	5	3	5
6	1	5	2	6	4	6	4	6
7	1	5	2	6	4	7	5	7
8	1	5	2	7	5	8	5	8
9	1	5	2	7	4	8	5	9
10			2	7	4	8	5	10

In table 1 the optimal dimensionality p^* for varying sample sizes n and starting dimensionalities ($p = 2 - 10$) for both removing the high variance directions (high) and low-variance directions (low) are shown. The left table shows the results when the eigenvalues are distributed between 1.5 and 0.666, in the right table the eigenvalues are between 2 and 0.5. The result for $n = 10, p = 10$ cannot be shown, because $n = 10$ objects can maximally span a 9-dimensional space. Note, that for high sample sizes it is in general bad to remove low-variance directions. Then the optimal dimensionality is always the original dimensionality. By removing high-variance directions, the performance stays constant, but the dimensionality can be reduced significantly. For small sample sizes and small differences in eigenvalues, the estimates of the low variance directions become inaccurate. In that case, the performance of retaining the high-variance directions is higher.

4 Experiments

This difference in performance can also be observed in real world datasets. Here we consider two datasets. The first is the face database [HTP00,Sun96], containing 19×19 gray value images of faces and non-faces. The training set consist of 2429 target objects (faces) and the testing set contains 472 face and 23573 non-face objects. The second dataset is the standard Concordia dataset [Cho96] containing 32×32 black-and-white images of handwritten digits. For each class, 400 objects were available for training and testing. In this experiment, objects of class '7' were considered target objects, all other objects are outliers. In figure 5

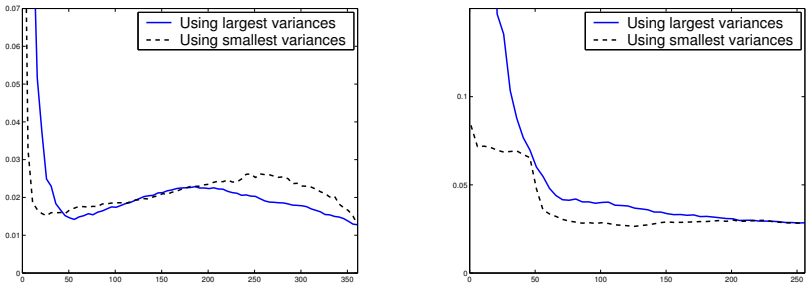

Fig. 5. The AUC of the face dataset (left) and the Concordia handwritten digits (right) for varying dimensionality, by removing the high and low variance directions.

the area under the ROC curve (AUC, a more robust error measure for one-class classification than standard classification error [Met78]) for varying dimensionalities is shown. When few dimensions are removed, using the high variance directions is often better, but at the very low dimensionalities, *removing* high-variance directions is best. This is according to the previous experiments.

5 Conclusions and discussion

We showed that reducing the dimensionality by PCA and retaining the high variance directions, is not always the best option for one-class classification. For high sample sizes and large differences in variances of the target distribution, retaining the low variance directions has smaller error. This is caused by the fact that the error on the outliers mainly depends on the eigenvalues of the estimated target covariance matrix S_n, and the error on the target class on the eigenvalues of $S_n \Sigma^{-1}$, where Σ is the true target covariance matrix.

Which approach is best depends on the basic bias-variance dilemma: how well can the different variance directions be estimated from a finite sample, and how well does this direction distinguish between the classes? These basic characteristics will be observed in other feature reduction methods, as Kernel PCA, Independent Component Analysis, Local Linear Embedding, etc.

Finally, it might seem that the OCC problem is now very simple. We can always transform the data, such that one, or a few directions have zero variance. This direction should then be used to get very good classification results (by equations (5) and (9)). Unfortunately, in practice the data cannot be scaled at will, because it is assumed that the outlier distribution is uniform. Transforming the data will therefore also transform the outlier distribution, and invalidate the results shown above.

Acknowledgements: This research was supported through a European Community Marie Curie Fellowship. The author is solely responsible for information communicated and the European Commission is not responsible for any views or results expressed. I would like to thank Klaus Müller for the useful discussions.

References

[And84] T.W. Anderson. *An introduction to multivariate statistical analysis.* John Wiley & Sons, 2nd edition, 1984.

[Cho96] Sung-Bae Cho. Recognition of unconstrained handwritten numerals by doubly self-organizing neural network. In *International Cconference on Pattern Recognition*, 1996.

[GBD92] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[Hes98] T. Heskes. Bias/variance decomposition for likelihood-based estimators. *Neural Computation*, 10:1425–1433, 1998.

[HR99] J. Hardin and D.M. Rocke. The distribution of robust distances. Technical report, University of California at Davis, 1999.

[HTP00] B. Heisele, Poggio. T., and M. Pontil. Face detection in still gray images. A.I. memo 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2000.

[JMG95] N Japkowicz, C. Myers, and M. Gluck. A novelty detection approach to classification. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 518–523, 1995.

[Jol86] I.T. Jollife. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[Met78] C.E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, VIII(4), October 1978.

[MH96] M.M. Moya and D.R. Hush. Network contraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[RGG95] G. Ritter, M.T. Gallegos, and K. Gaggermeier. Automatic context-sensitive karyotyping of human elliptical symmetric statistical distributions. *Pattern Recognition*, 28(6):823–831, December 1995.

[SPST+99] B Schölkopf, J. Platt, J. Shawe-Taylor, Smola A., and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 1999.

[Sun96] K.-K. Sung. *Learning and Example Selection for Object and Pattern Recognition.* PhD thesis, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA, 1996.

[Tax01] D.M.J. Tax. *One-class classification.* PhD thesis, Delft University of Technology, http://www.ph.tn.tudelft.nl/~davidt/thesis.pdf, June 2001.

[Wil62] S. Wilks. *Mathematical statistics.* John Wiley, 1962.