One-Class Classification with Subgaussians

Amir Madany Mamlouk[1], Jan T. Kim[1], Erhardt Barth[1], Michael Brauckmann[2], and Thomas Martinetz[1]

[1] Institute for Neuro- and Bioinformatics, University of Lübeck
[2] ZN Vision Technologies AG, Bochum
{madany,kim,barth,martinetz}@inb.uni-luebeck.de

Abstract. If a simple and fast solution for one-class classification is required, the most common approach is to assume a Gaussian distribution for the patterns of the single class. Bayesian classification then leads to a simple template matching. In this paper we show for two very different applications that the classification performance can be improved significantly if a more uniform subgaussian instead of a Gaussian class distribution is assumed. One application is face detection, the other is the detection of transcription factor binding sites on a genome. As for the Gaussian, the distance from a template, i.e., the distribution center, determines a pattern's class assignment. However, depending on the distribution assumed, maximum likelihood learning leads to different templates from the training data. These new templates lead to significant improvements of the classification performance.

1 Introduction

In many applications a one-class classification problem has to be solved, i.e. the separation of a single class of patterns from the rest of the pattern space. A typical example is the detection of faces in images: the class of human faces has to be separated from all the other possible patterns [3, 6, 12]. Usually, the single class occupies only a negligible volume compared the rest of the pattern space, and only positive examples for this class are useful or even given for training a classifier. The huge rest of the pattern space can hardly be represented by examples. Another typical one-class classification problem can be found in bioinformatics. Gene regulation is controlled by sequence-specific DNA binding proteins, the so-called transcription factors [2]. Molecular biologists want to know the sites where these factors bind on a genome. Sequence patterns where binding takes place form the single class that has to be separated from the rest of all possible sequence patterns.

Our investigation is further motivated by the computational requirements for an industrial face detection system that has to find faces in a video stream in real time. A common and simple approach is to assume a Gaussian distribution for the single class. More complex would be a mixture of Gaussians or the application of a support vector machine [6, 10]. Therefore simple approaches are needed, but what performance can be achieved by simple template matching? As we will see below, Bayesian classification with a Gaussian class distribution leads to template matching. However, the same is true for every radial symmetric

In: B.Michaelis and G.Krell (Eds.): DAGM 2003, LNCS 2781, pp. 346–353, 2003.

and monotonic class distribution, with the form of the distribution determined only by the template. We focused our investigation on a class of subgaussian distributions that vary from the Gaussian to the rectangular distribution. Obviously, the subgaussian model which fits most adequately to the distribution of the given data will yield the best template.

2 One-class classification with Gaussians

We assume patterns $\mathbf{x} \in \mathbb{R}^N$. Given a pattern $\mathbf{x}$, we ask whether this pattern belongs to the class c. For example, whether an image pattern is a face. Bayes decision rule answers this based on the posterior class probability $P(c|\mathbf{x})$ of the given pattern to belong to class c. $P(c|\mathbf{x})$ can be derived from the pattern distribution of class c, $P(\mathbf{x}|c)$, with the Bayes theorem $P(c|\mathbf{x})P(\mathbf{x}) = P(\mathbf{x}|c)P(c)$. $P(c)$ is the prior class probability, and $P(\mathbf{x})$ denotes the prior probability for the occurrence of pattern $\mathbf{x}$.

A simple model for the probability distribution $P(\mathbf{x}|c)$ of patterns from class c is the Gaussian. Assuming the same prior probability $P(\mathbf{x})$ for all patterns, we obtain for the posterior class probability

$$P(c|\mathbf{x}) = Ce^{-(\mathbf{x}-\mathbf{w})^2/2\sigma^2} \tag{1}$$

with $\mathbf{w}$ as the center of the Gaussian, σ^2 as its variance, and C as a normalization constant that includes the prior class probability $P(c)$.

We assume that a query pattern $\mathbf{x}$ belongs to class c, if $P(c|\mathbf{x})$ exceeds a prespecified value $P_{\min}$. Hence, all the patterns that lie within a sphere with a certain radius R around $\mathbf{w}$ are assumed to belong to class c. Thus, the distribution center $\mathbf{w}$ can be regarded as a template for the class c.

Usually, the prior class probability $P(c)$ is not known, and, hence, a prespecified $P_{\min}$ can not explicitly be translated into a respective radius R. But one knows that $P_{\min}$ increases with decreasing R. By varying R one controls the specificity/sensitivity of the classifier.

Starting with $R = 0$ and increasing R against infinity, one obtains the so-called specificity-sensitivity- or receiver-operating-curve (ROC) of the classifier.

2.1 Determining the template by maximum likelihood estimation

To estimate the template, i.e., the center of the Gaussian distribution, based on example patterns $X = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ for class c (the training set), we assume that the training patterns are independently drawn from the class distribution $P(\mathbf{x}|c)$. Then the likelihood $L(X|\mathbf{w})$ that a Gaussian class distribution with center $\mathbf{w}$ generates the data X is given by

$$L(X|\mathbf{w}) = \prod_{\mathbf{x}\in X} e^{-(\mathbf{x}-\mathbf{w})^2/2\sigma^2}. \tag{2}$$

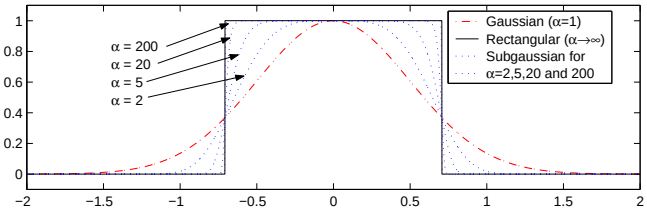

Fig. 1. Shape of the distribution functions $P_\alpha(\mathbf{x}|c)$ with $\mathbf{w} = 0$ and $\sigma = 0.5$ for $\alpha = 1$, which is the Gaussian (dash-dotted line), $\alpha = 2, 5, 20, 200$ (dotted lines), and for $\alpha = \infty$, which is the rectangular distribution (solid line).

Maximum likelihood estimation (MLE) looks for the $\mathbf{w}$ that maximizes this likelihood and yields as solution

$$\mathbf{w}^* = \frac{1}{p} \sum_{\mathbf{x} \in X} \mathbf{x}. \tag{3}$$

Given a Gaussian distribution, the center-of-gravity of the training data is the MLE for its center, i.e., one takes the average pattern as a template.

3 One-class classification with subgaussians

Instead of being a Gaussian, we now assume the distribution of the patterns of class c to be a subgaussian

$$P(\mathbf{x}|c) = C_\alpha e^{-[(\mathbf{x}-\mathbf{w})^2/2\sigma^2]^\alpha}. \tag{4}$$

For $\alpha = 1$ we obtain the standard Gaussian distribution discussed above. Now we are interested in α-values larger than one. Fig. 1 illustrates how the shape of the distribution $P(\mathbf{x}|c)$ changes with an increasing α towards a more and more rectangular (spherical in higher dimensions) distribution. For $\alpha \to \infty$ the distribution $P(\mathbf{x}|c)$ becomes a (hyper)sphere with $P(\mathbf{x}|c) = 1$ inside the sphere, i.e., for $(\mathbf{x} - \mathbf{w})^2 < 2\sigma^2$, and zero outside.

As above, with the assumption of a homogeneous prior $P(\mathbf{x})$ the posterior class probability $P(c|\mathbf{x})$ has the same shape as the class pattern distribution $P(\mathbf{x}|c)$. Again, all the patterns inside a hypersphere of radius R and center $\mathbf{w}$, are assigned to class c. By varying R, the false acceptance rate (FAR) and the false rejection rate (FRR) can be influenced along the respective ROC. Compared to the standard Gaussian distribution, we now obtain different distribution centers $\mathbf{w}$, i.e., different templates. Should the real pattern distribution rather have the shape of a sphere than of a Gaussian, the classification performance will improve with increasing α.

3.1 Determining the center of the subgaussian distribution

As for the Gaussian distribution we determine the center of the subgaussian by MLE. The likelihood $L(X|\mathbf{w})$ of generating the data X is

Fig. 2. Face detection templates. From left to right the average face ($\alpha = 1$) and the subgaussian templates are shown for $\alpha = 10, 20, 100$ and $\alpha \to \infty$.

$$L(X|\mathbf{w}) = \prod_{\mathbf{x} \in X} e^{-[(\mathbf{x}-\mathbf{w})^2/2\sigma^2]^\alpha}. \tag{5}$$

Maximizing $L(X|\mathbf{w})$ is equivalent to minimizing

$$S(\mathbf{w}) = \sum_{\mathbf{x} \in X} (\mathbf{x} - \mathbf{w})^{2\alpha}. \tag{6}$$

The maximization of the likelihood with respect to $\mathbf{w}$ is independent of the σ that is assumed for the model distribution. If we set the derivative of $S(\mathbf{w})$ to zero, we obtain an equation for the distribution center $\mathbf{w}^*_\alpha$. However, this equation can be solved explicitly only for $\alpha = 1$. In this case the solution is, as expected, the center-of-gravity of the training data. For $\alpha > 1$, $\mathbf{w}^*_\alpha$ can be obtained iteratively by gradient descent. The new estimate $\mathbf{w}^*_\alpha(t+1)$ then follows from the old estimate according to

$$\mathbf{w}^*_\alpha(t+1) = \mathbf{w}^*_\alpha(t) + \epsilon \sum_{\mathbf{x} \in X} ||\mathbf{x} - \mathbf{w}^*_\alpha(t)||^{2(\alpha-1)} (\mathbf{x} - \mathbf{w}^*_\alpha(t)), \tag{7}$$

with ϵ as the step size of the gradient descent.

4 Face Detection

As a first example we present the results we obtained for face detection. For these experiments we used a public database[3] of gray-scale images with (19×19) pixel resolution. This database is a standard for the evaluation of face detection systems [3, 6, 8, 10], and thus, is a good basis for comparisons. The database provides 2,429 faces as a training set, and 472 faces and 23,573 non-faces as a test set. We flipped the training images horizontally to increase the training set to 4,858 faces. As in [8], the pixel intensities of each image were normalized to zero-mean and unit variance.

With the training faces the templates $\mathbf{w}^*_\alpha$ were calculated for different α. For $\alpha = 1$ we simply had to calculate the average face. Starting from this average face, α was increased step by step. For each α, 100 iterations of the gradient descent ($\epsilon = 0.01$) were performed. In Fig. 2 we see how the template $\mathbf{w}^*_\alpha$ changed with increasing α.

[3] http://www.ai.mit.edu/projects/cbcl/software-datasets/FaceData2.html

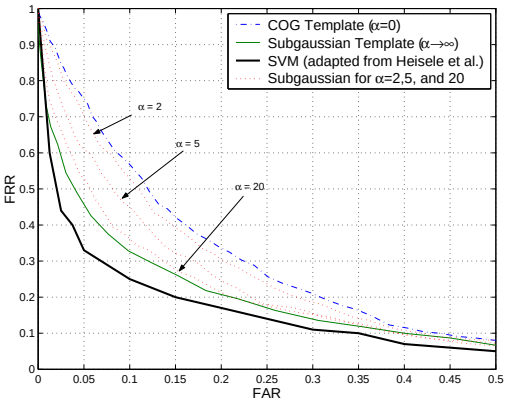

Fig. 3. Face detection performance for different templates. The dash-dotted line shows the ROC for the average face as a template ($\alpha = 1$). The small solid line shows the ROC for the template one obtains if a uniform spherical pattern distribution is assumed ($\alpha \to \infty$). The performance increases with increasing α. The thick solid line is the ROC for a SVM classifier adapted from Heisele et al. [3].

Figure 3 presents the ROC for different templates. The closer such a curve bends towards the origin, the better the classifier works. Interestingly, the performance increases significantly with increasing α. For the common Gaussian ($\alpha = 1$) we obtain the worst performance. The best performance is obtained for $\alpha \to \infty$, which yields the center of the spherical distribution as template. Obviously, the distribution of faces in the pixel space is rather spherical than Gaussian.

To illustrate the gain in performance in a more global context, we adapted a ROC obtained with a state-of-the-art approach for face detection by Heisele et al. [3]. A support vector machine (SVM) with a 2nd degree polynomial kernel was trained not only on the 2,429 faces that we used, but in addition also on 4,450 non-face images. As a preprocessing step, the histogram of each image was equalized. Interestingly, with our extremely simple approach of template matching within the raw pixel space, we obtain a classification performance which is remarkably close to this much more sophisticated approach of Heisele et al. (Fig. 3).

5 Detection of protein-DNA binding sites

Genetic information in most biological systems is stored in DNA sequences, consisting of base pairs which are denoted by A (adenine), C (cytosine), G (guanine) and T (thymine). Regulation of gene expression is mediated by specialized proteins, called transcription factors, which bind to specific regulatory sites on the genome more tightly than to all other sites. A transcription factor "recognizes" a binding site by the local sequence, called the binding word.

For understanding and modelling gene regulation, it is necessary to know at which words a certain transcription factor binds and executes its function. Classification of words into binding words and non-binding words is therefore an important task in bioinformatics. Since for experimental reasons non-binding words have only rarely been described, we have to solve a one-class classification problem in which only positive samples from the class of binding words are given.

For representing DNA sequence information by vectors with real-valued components, a method called "orthogonal coding" is typically used. Let $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ denote the alphabet of base pairs. Orthogonal coding is a mapping from $\mathcal{A}^L$ to $\mathbb{R}^{4L}$ which represents a sequence of L nucleotide symbols as a vector $\mathbf{x} = (x_{\mathrm{A},1}, x_{\mathrm{C},1}, x_{\mathrm{G},1}, x_{\mathrm{T},1}, x_{\mathrm{A},2}, \ldots, x_{\mathrm{T},L})$, where $x_{b,l} = 1$ if the l-th symbol is b and $x_{b,l} = 0$ otherwise. As an example, the orthogonal coding of the sequence `GAT` is $(0,0,1,0,1,0,0,0,0,0,0,1)$. By this construction, symbols are represented by quartets of components, and different symbols correspond to different, orthogonal quartets.

5.1 Classification of words

The concept presented above to choose a template and to use the distance to the template as a basis for classification is now applied to the problem of detecting binding words. A word (pattern) $\mathbf{x}$ is assumed to be a binding word if the distance $||\mathbf{x} - \mathbf{w}||$ of the word to the template $\mathbf{w}$ is smaller than a maximum distance R. Since $||\mathbf{x}||^2 = L$ is valid within the orthogonal coding scheme, we obtain $||\mathbf{x} - \mathbf{w}||^2 = L + ||\mathbf{w}||^2 - 2\mathbf{w}^T\mathbf{x}$. Hence, the condition of a maximum distance R then is equivalent to the condition that the so-called score $\mathbf{w}^T\mathbf{x}$ of a word has to exceed a threshold $S_{\min}$ for being a binding word.

The template $\mathbf{w}$ is called a "scoring matrix" in the bioinformatics literature, where the term "matrix" refers to the common practice to arrange the $4L$ components of $\mathbf{w}$ in a $4 \times L$ table. The scoring matrix is a good approximation for calculating the protein-DNA binding energy [1, 5] and, therefore, it is structurally adequate for capturing the binding behaviour of a transcription factor. The maximum likelihood template now depends on the assumed distribution of the binding words.

Here we assume the subgaussian class distributions as introduced in Section 3. For $\alpha = 1$, the case of a Gaussian class distribution, we obtain for the template $\mathbf{w}$ the arithmetic mean of the experimentally known binding words. Interestingly, this is equivalent to the so-called profile matrix, the most commonly used approach for binding word detection in bioinformatics [2, 9].

As an alternative, we now consider the template resulting from $\alpha \to \infty$ and ask, whether this yields improvements similar to those we have seen for face detection. This template is equivalent to the so-called binding matrix which has been introduced recently [5].

5.2 Sample application

Since non-binding words are not available, it is not possible to determine ROCs as for the face detection problem. Instead, we split up the known binding words into a training set and a test set. The training set is then used to calculate the templates for $\alpha = 1$ and $\alpha \to \infty$. The respective thresholds are then set to the largest value that still provides a zero FRR on both the test set and the training set (i.e. we require that all binding words are classified correctly). The corresponding FAR cannot directly be determined, because false positives are not known. However, from theoretical analyses and empirical observation it is

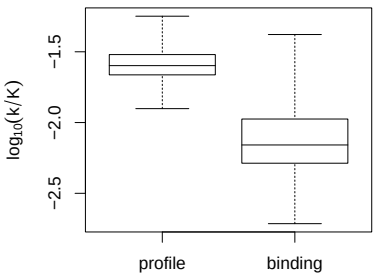

Fig. 4. Box plot showing the gain in specificity achieved with the assumption of a subgaussian instead of a Gaussian distribution. From the 73 binding words for the SOX-9 transcription factor, 1000 subsets of 48 binding words were independently drawn at random and k/K was determined as described in the text. Boxes encompass the middle quartiles, the bar depicts the median and the whiskers extend to the extreme values.

known that the ratio of k, the number of words accepted by the transcription factor, to $K = 4^L$, the number of all words, has to be small [7, 4]. Quantitatively, $k/K \approx 10^{-3}$ is, on average, a reasonable estimate and $k/K < 10^{-2}$ can definitely be expected to be satisfied. Thus, the amount by which the k/K ratio obtained for the classifier exceeds 10^{-3} is a measure for the FAR which corresponds to the zero FRR.

Of course, it would be easy to achieve a low k/K at FRR=0, if we could choose a classifier of arbitrary structure. However, we want to minimize the k/K ratio with a template (matrix) classifier since this approach is structurally adequate from biochemical and biophysical considerations. As we will see, within this approach it is not easy to achieve a k/K ratio which is in the right order of magnitude. In particular, the profile matrix typically leads to a k/K ratio which indicates a large FAR. Finding a template (matrix) that allows for a lower k/K value would therefore be a significant improvement.

Fig. 4 shows the results we achieved for $\alpha = 1$ and $\alpha \to \infty$ on a set of 73 binding words of the SOX-9 transcription factor provided by the TRANSFAC database [11] (TRANSFAC matrix `M00410`). We created training sets by randomly drawing 2/3 of the binding words. Templates were computed based on these training sets, and for each template, the threshold was set to the maximum value at which all known binding words, including those not used for training, are classified correctly. Then the corresponding k/K ratios were determined. For an extreme subgaussian distribution ($\alpha \to \infty$) we achieve a reduction of k/K by half an order of magnitude in comparison to the classifier based on the Gaussian distribution ($\alpha = 1$), which corresponds to the common profile matrix.

6 Conclusions

We have shown that the assumption of a subgaussian class distribution can increase one-class classification performance significantly compared to the widespread assumption of a Gaussian class distribution. This is obtained with simple and fast template matching as in case of the traditional Bayesian approach for Gaussian class distributions. It is just that the template that results from assuming the subgaussian class distribution seems to be more appropriate. With

the framework presented in this paper, the shape of the distribution is steered by a single continuous parameter α, with $\alpha = 1$ for the standard Gaussian and $\alpha \to \infty$ for a rectangular distribution. The corresponding templates can be determined by simple gradient descent. The performance increase was shown for two example problems as different as face finding and DNA-binding site detection. For face detection, the increase was robust against different methods of image preprocessing like histogram normalization, subtraction of best-fit linear plane, and edge extraction. This suggests that there might be quite a number situations and problems where the assumption of a subgaussian instead of a Gaussian pattern distribution would be more appropriate and thus lead to superior results without additional computational costs for the classification task.

Acknowledgements

The authors would like to thank Karsten Hinckfuß and Frank Prill for preparing the data for the face detection analysis as well as Jan Gewehr for his preparation of the TRANSFAC data. This work has been supported by the MORPHA project of the BMBF, FKZ01IL902Q/0.

References

1. P. V. Benos, M. L. Bulyk, and G. D. Stormo. Additivity in protein-DNA interactions *Nucleic Acids Research*, 30:4442–4451, 2002.
2. K. Frech, K. Quandt, and T. Werner. Finding protein-binding sites in DNA sequences: The next generation. *TIBS*, 22:103–104, 1997.
3. B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. Technical Report AI Memo 1687, Massachusetts Institute of Technology, 2000.
4. J. T. Kim, T. Martinetz, and D. Polani. Bioinformatic principles underlying the information content of transcription factor binding sites. *Journal of Theoretical Biology*, 220:529–544, 2003.
5. T. Martinetz, J. E. Gewehr, and J. T. Kim. Statistical learning for detecting protein-DNA-binding sites. In *Int. Joint Conf. on Neural Networks 2003*. IEEE Press, 2003.
6. E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. *Proceedings of CVPR'97*, 1997.
7. T. D. Schneider, G. D. Stormo, and L. Gold. Information content of binding sites on nucleotide sequences. *J.Mol.Biol.*, 188:415–431, 1986.
8. H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. *Proceedings of CVPR'98*, 1998.
9. G. D. Stormo. DNA binding sites: Representation and discovery. *Bioinformatics*, 16:16–23, 2000.
10. K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE PAMI*, 20:39–51, 1998.
11. E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüß, I. Reuter, and F. Schacherer. TRANSFAC: An integrated system for gene expression regulation. *Nucl. Acids Res.*, 28:316–319, 2000.
12. M-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE PAMI*, 24:34–58, 2002.