8 A New Perspective: Working on the Distance Measure

Speaker verification is a highly unbalanced two-class classification problem and it might be important to consider specific training criteria for such cases. Gradient based models (such as Multilayer Perceptrons) can easily accommodate various possible training criteria adapted to unbalanced datasets, and thus can be good candidates to solve this problem. Unfortunately, when using a large margin approach, the number of training iterations needed to converge to a good solution is huge. This has also been observed in Collobert and Bengio (2004). SVMs have usually faster convergence rates, so we will instead consider unbalanced criteria for SVMs.

After analyzing two already proposed criteria for this problem (Lin et al., 2002) and (Grandvalet et al., 2005), we note that they are useless in our case. Indeed, empirically we observed that for all SVM based sequence kernels that give reasonable performance, and for all client models, the problem is in fact linearly separable in the feature space and we can show that for such a problem these unbalanced criteria have no effect. Moreover, in the separable case the standard SVM solution is good because only examples in the margin are considered.

At the opposite, another specific speaker verification problem, which for us is more important, is addressed here: the intra-impostor distance distribution is different than the intra-client distance distribution. We thus propose to modify

R. Collobert and S. Bengio. Links between perceptrons, MLPs and SVMs. In *International Conference on Machine Learning, ICML*, 2004.

Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

Y. Grandvalet, J. Mariéthoz, and S. Bengio. A probabilistic interpretation of svms with an application to unbalanced classification. In *Advances in Neural Information Processing Systems, NIPS 15*, 2005. IDIAP-RR 05-26.

the SVM kernel by assuming a Gaussian noise on negative examples. Starting from a principled approach, and after some empirical modification, we show that the new system outperforms the baseline system.

The outline of this chapter goes as follows. In Section 8.1, we present the known unbalanced class criteria for SVMs and show they are useless for separable problems. Section 8.2 is dedicated to a new similarity measure that takes into account the difference between the intra-impostor and intra-client distance distributions.

8.1 Unbalanced SVM Criteria

SVMs are known to perform well in terms of misclassification error, but they also have been recognized to provide skewed decision boundaries for unbalanced classification losses, where the losses associated with incorrect decisions differ according to the true label. The mainstream approach used to address this problem was proposed in (Lin et al., 2002) and consists in using different costs for positive and negative examples using two smoothing parameters C_+, C_- instead of a single C as in (2.9). This solution was used, for instance, in Chapter 6 and is given in (6.5).

Another solution, proposed in Grandvalet et al. (2005) is based on a probabilistic interpretation of SVMs. The cost to optimize now becomes:

$$\begin{aligned} \arg\min_{(\mathbf{w},b)} \frac{\| \mathbf{w} \|^2}{2} \quad + \quad C \Bigg(& \sum_{\{i|y_i=1\}} [-\log(P_0) - (1-P_0)(f(\mathbf{x}_i)+b)]_+ \\ & + \sum_{\{i|y_i=-1\}} [-\log(1-P_0) + P_0(f(\mathbf{x}_i)+b)]_+ \Bigg) \end{aligned} \tag{8.1}$$

where $P_0 = \frac{C(\mathrm{FP})}{C(\mathrm{FP})+C(\mathrm{FN})}$, $C(\mathrm{FP})$ is the cost of a false positive and $C(\mathrm{FN})$ is the cost of a false negative.

Even if these two approaches give good results on standard machine learning databases, as shown in (Grandvalet et al., 2005), they have no positive effect in our case. Indeed, empirically we can observe that for all sequence kernels that provide good performance, the problem is separable: all the training examples are well classified. It seems reasonable: the feature space dimension is greater

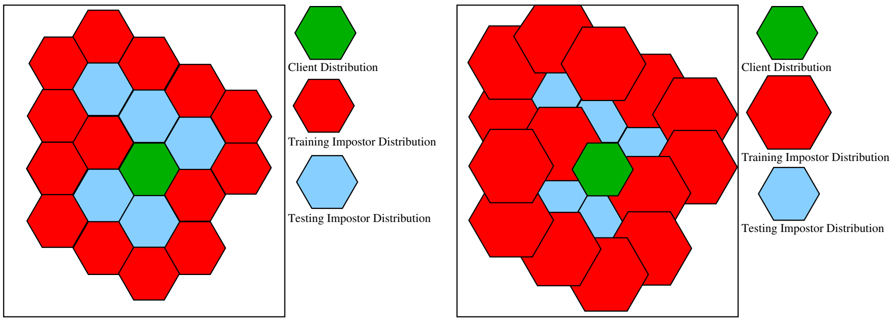
Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in non-standard situations. *Machine Learning*, 46:191–202, 2002.

Y. Grandvalet, J. Mariéthoz, and S. Bengio. A probabilistic interpretation of svms with an application to unbalanced classification. In *Advances in Neural Information Processing Systems, NIPS 15*, 2005. IDIAP-RR 05-26.

than the number of training examples. Moreover most of the time the optimal value for C tends to ∞ and thus the criterion does not tolerate any error. This is probably because it cannot make an error on positive examples: they are too few; and it can neither tolerate an error on a negative example: the coverage of the training negative examples is not good enough. Indeed, each negative example can cover its own variability but cannot cover the future testing negative examples (other impostors). As the training positive and negative examples do not correspond well enough to the test set, it can be interesting to use prior knowledge in the kernel: for instance we expect the variance of the intra-impostor distance distribution to be larger than that of the intra-client distance distribution.

8.2 Class Dependent RBF Kernel

When a two-class classification problem is separable, we can admit that a solution maximizing the margin is a good idea even if the problem is unbalanced. Indeed an SVM considers only examples in the margin and ignores other examples. Hence, the standard SVM criterion can be good also for separable unbalanced class problems. It still remains that, in the case of speaker verification, the distribution of the distance between two impostor accesses is larger than the client distance distribution: impostors are individual speakers and thus the intra-impostor distribution is more similar to the inter-class distance distribution than the intra-client distribution. In this case, it can be a good idea to change the kernel in order to make the negative examples closer. In other words, a negative example should cover its own variability (same speaker), but also unseen negative examples (other impostors).


Client Distribution
Training Impostor Distribution
Testing Impostor Distribution
Client Distribution
Training Impostor Distribution
Testing Impostor Distribution

(a) Normal

(b) Enlarged

Figure 8.1. Client, training and testing impostor distributions.

Figure 8.1 shows that enlarging the negative example distribution, for instance by using a larger σ value for intra-negative RBF kernel evaluation, increases the coverage of the unseen impostor examples.

Vapnik (2000) proposed the use of vicinal risk minimization to learn a decision function over distributions instead of points. One of several solutions he proposed is the soft vicinity function that uses a kernel over distributions. The main idea is to assume a Gaussian noise over each negative example. Using an RBF kernel with a Gaussian noise distribution, we have:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2\pi\sigma_i\sigma_j} \int\int \exp\left\{-\frac{(\mathbf{x}-\mathbf{x}')^2}{2\sigma^2} - \frac{(\mathbf{x}'-\mathbf{x}_i)^2}{2\sigma_i^2} - \frac{(\mathbf{x}-\mathbf{x}_j)^2}{2\sigma_j^2}\right\} d\mathbf{x}\, d\mathbf{x}' \tag{8.2}$$

where σ is the RBF kernel hyper-parameter, σ_i the noise standard deviation of example $\mathbf{x}_i$ and σ_j the noise standard deviation of example $\mathbf{x}_j$.

Vapnik (2000) then showed that (8.2) can be rewritten as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \frac{\sigma_i^2}{\sigma^2} + \frac{\sigma_j^2}{\sigma^2}\right)^{(-\frac{d}{2})} \exp\left\{-\frac{(\mathbf{x}_i - \mathbf{x}_j)}{2(\sigma^2 + \sigma_i^2 + \sigma_j^2)}\right\} \tag{8.3}$$

where d is the dimension of the input vector.

Let us now consider a Gaussian noise for the negative examples only, with variance $\tau\sigma^2$ where τ is a constant to tune, we obtain:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \exp -\frac{(\mathbf{x}_i-\mathbf{x}_j)^2}{2\,\sigma^2} & \text{if } y_i = y_j = 1 \\ (1+\tau)^{(-\frac{d}{2})} \exp -\frac{(\mathbf{x}_i-\mathbf{x}_j)^2}{2\,\sigma^2(1+\tau)} & \text{if } y_i \neq y_j \\ (1+2\tau)^{(-\frac{d}{2})} \exp -\frac{(\mathbf{x}_i-\mathbf{x}_j)^2}{2\,\sigma^2(1+2\,\tau)} & \text{if } y_i = y_j = -1. \end{cases} \tag{8.4}$$

In (8.4) we have a kind of RBF kernel with larger standard deviation if $y_i = y_j = -1$ than otherwise. This is what we expected: make the intra-negative distance smaller. Unfortunately, the constant $(1+2\tau)^{(-\frac{d}{2})}$ has the inverse effect and decreases faster that the exponential term. Moreover Vapnik (2000) said nothing about how to choose σ for a new test point (for which the class is obviously not known).

Even if this is not principled, we would like to propose some simplifications to Vapnik's approach, as follows:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \exp -\frac{(\mathbf{x}_i-\mathbf{x}_j)^2}{\sigma_{++}^2} & \text{if } y_i = y_j = 1 \\ \exp -\frac{(\mathbf{x}_i-\mathbf{x}_j)^2}{\sigma_{+-}^2} & \text{if } y_i \neq y_j \\ \exp -\frac{(\mathbf{x}_i-\mathbf{x}_j)^2}{\sigma_{--}^2} & \text{if } y_i = y_j = -1 \end{cases} \tag{8.5}$$

REF V. N. Vapnik. *The nature of statistical learning theory.* Springer, second edition, 2000.

with

$$\sigma_{++} = \sigma{+-} = \sigma_{+} \tag{8.6}$$

$$\sigma_{--} = \sigma_{-} \tag{8.7}$$

$$\sigma_{-} > \sigma_{+} \tag{8.8}$$

where σ_- and σ_+ are hyper-parameters to tune. The differences between (8.4) and (8.5) are that we remove the constants involving the dimension of the data d, and choose the same value for σ_{++} and σ_{+-}; in fact when we have only one positive example to train the model, any value for σ_{++} yields the same kernel value (equal to one). During test, we tried empirically several values of σ between σ_+ and σ_- and found that the best value is σ_+ for both Banca and Polyvar databases.

Figure 8.2 shows that the vicinity based method outperforms the Max operator based RBF kernel on the development set of the Polyvar database. This is also confirmed on the test set on Figure 8.3 and Table 8.1. The two models are statistically significantly different for most value of γ.

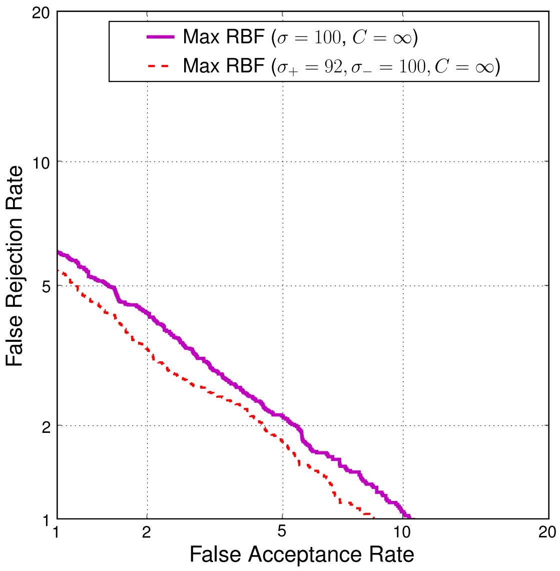


Figure 8.2. DET curves on the development set of the Polyvar database for the best σ, σ_+, σ_- Max RBF kernel.

We also performed the same experiments on the Banca database and draw the same conclusion as shown in Figure 8.4, Figure 8.5 and Table 8.2. Even

Table 8.1. Results on the test set of the Polyvar database for Mean and Max operators for polynomial and RBF σ and σ_+, σ_- kernels.

	GMM $N_g = 100$	Mean $p = 3$ $C = \infty$	Max $\sigma = 100$ $C = \infty$	Max $\sigma_+ = 92$ $\sigma_- = 100$ $C = \infty$
HTER [%]	4.9	4.47	4.21	3.59
95% Confidence	±0.34	±0.32	±0.28	±0.32
# Support Vectors	-	87	99	76

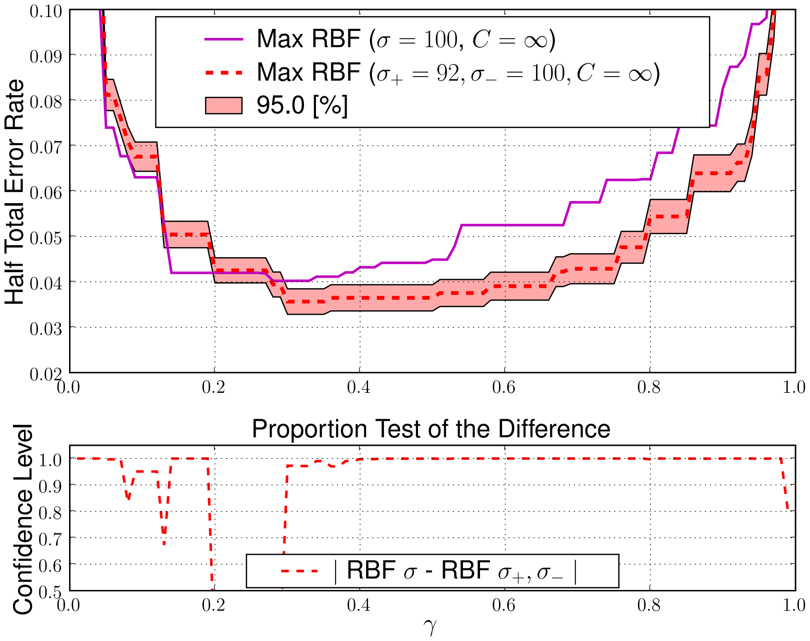


Figure 8.3. EPC curves on the test set of the Polyvar database for the best σ, σ_+, σ_- Max RBF kernel.

if on this database the results are not statistically significantly different due to the size of this database, the effect seems positive. Note also that, for this database, we are still far from the GMM based system, on the test set as seen in Table 8.2 and Figure 8.5 but it seems not be the case on the development set as seen in Figure 8.4.

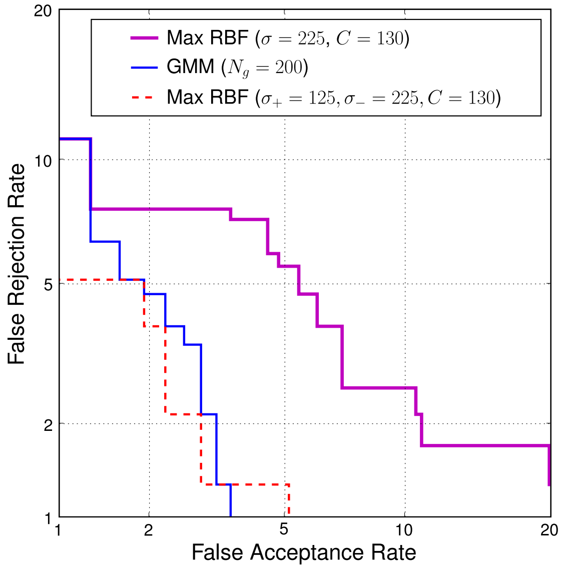


Figure 8.4. DET curves on the development set of the Banca database for the best σ, σ_+, σ_- Max RBF kernel and GMM based system.

Table 8.2. Results on the test set of the Banca database for Mean and Max operators for polynomial and RBF σ and σ_+, σ_- kernels.

	GMM $N_g = 200$	Mean $p = 3$ $C = \infty$	Max $\sigma = 225$ $C = 130$	Max $\sigma_+ = 125$ $\sigma_- = 225$ $C = 130$
HTER [%]	2.72	6.57	4.7	4.11
95% Confidence	±1.42	±2.1	±1.78	±1.66
# Support Vectors	-	27	17	13

8.3 Conclusion

In this chapter, we considered the unbalanced class problem underlying the speaker verification task. We tried to use modified criteria for SVM in order to deal with unbalanced datasets and observed that they have no effect on separable problems, which is the case for our speaker verification experiments. Indeed, we enlight the fact that for separable problems, the standard SVM criterion gives a good solution even with highly unbalanced task.

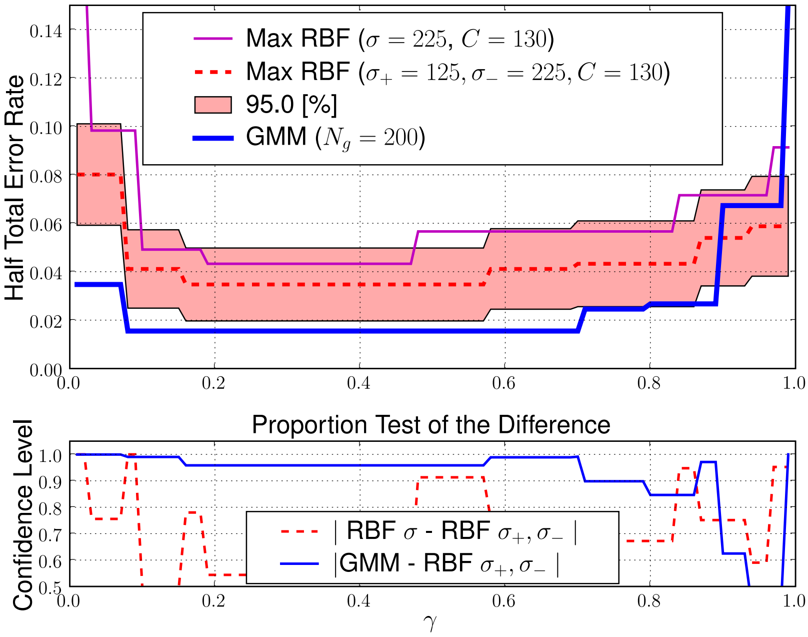


Figure 8.5. EPC curves on the test set of the Banca database for the best σ, σ_+, σ_- Max RBF kernel and GMM based system.

We proposed, instead, to work on new similarity measures. The intra-impostor distance distribution is larger than the intra-client distribution due to the problem itself. We thus proposed, based on the idea of the vicinity function proposed by Vapnik (2000), to add a Gaussian noise over the negative examples only. Unfortunately, we had to apply some empirical simplification in order to make this new approach feasible, which made it less principled. However, this suggests to modify the standard similarity measure, for example by adapting the kernel (Kwok and Tsang, 2003) or by learning a similarity measure, as done by Chopra et al. (2005) for face verification.

V. N. Vapnik. *The nature of statistical learning theory.* Springer, second edition, 2000.

J. Kwok and I. Tsang. Learning with idealized kernels. In *International Conference on Machine Learning, ICML*, 2003.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.