Application of Maximum Distance Minimization to Gene Expression Data

Jens Hocke, Thomas Martinetz

Institute for Neuro- and Bioinformatics, University of Lübeck

1 Introduction

The k-Nearest-Neighbor (k-NN) [1] algorithm is a popular non-linear classifier. It is simple and easy to interpret. However, the often used Euclidean distance is an arbitrary choice, because the data dimensions are not scaled according to their relevance. Similar to relevance learning in the context of LVQ classifiers [2], the scaling of the dimensions can be adapted by feature weighting to improve the classification rate of k-NN.

An optimal rescaling has to minimize the classification error $E(X)$ of the k-NN algorithm. Often this problem is called the *feature weighting* problem. We want to find a weight vector $\boldsymbol{w} \in \Re^D, w_\mu \geq 0, \mu = 1, ..., D$ for some given dataset $X = \{\boldsymbol{x}_i \in \Re^D, i = 1, ..., N\}$ that helps the classifier to minimize $E(X)$. In case the Euclidean distance is used, the weighted distance between two data points $\boldsymbol{x}, \boldsymbol{x}'$ becomes $d(\boldsymbol{x}, \boldsymbol{x}') = ||\boldsymbol{x} - \boldsymbol{x}'||_{\boldsymbol{w}} = \sqrt{\sum_{\mu=1}^{D} w_\mu (x_\mu - x'_\mu)^2}$.

Well known methods for feature weighting are Relief [3] and Simba [4]. Related is the more general problem of *metric learning* with Large Margin Nearest Neighbor Classification (LMNN) [5] as a popular approach, that optimizes the Mahalanobis distance $d(\boldsymbol{x}, \boldsymbol{x}') = ||\boldsymbol{x} - \boldsymbol{x}'||_W = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^T W (\boldsymbol{x} - \boldsymbol{x}')}$.

We here present a method that contrary to the other methods is independent of the initial dimension scaling and evaluate it on gene expression data.

2 Maximum Distance Minimization

For rescaling the dimensions, we do not look at local neighbors, as the other methods do. Instead we try to minimize, by a very global optimization, the maximum distance between all pairs of data points of the same class, while keeping the pairwise distance between data points of different classes large. We therefore name our method Maximum Distance Minimization (MDM). Formally, we are solving the following constrained optimization problem

$$||\boldsymbol{x}_i - \boldsymbol{x}_l||^2_{\boldsymbol{w}} \geq 1 \; \forall i, l : y_i \neq y_l \tag{1}$$

$$||\boldsymbol{x}_i - \boldsymbol{x}_j||^2_{\boldsymbol{w}} \leq r \; \forall i, j : y_i = y_j \tag{2}$$

$$\min_{\boldsymbol{w}} r \quad w_\mu \geq 0 \; \forall \mu, \tag{3}$$

where y_i, y_l, and y_j are the class labels of $\boldsymbol{x}_i$, $\boldsymbol{x}_l$, and $\boldsymbol{x}_j$. The above problem can be formulated as a linear program, which is always solvable, even without slack variables.

	Euclidean	MDM	Relief	Simba	LMNN
Breast Cancer	**8.07(6.13)**	11.42(7.25)	9.68(7.09)	14.07(7.63)	9.78(7.13)
	1213.00(0.00)	364.76(62.65)	1213.00(0.00)	1213.00(0.00)	1136.96(0.75)
DLBCL	13.11(5.24)	14.67(5.33)	**11.17(5.03)**	13.56(6.06)	15.44(4.32)
	661.00(0.00)	293.86(34.13)	661.00(0.00)	661.00(0.00)	559.54(1.99)
Leukemia	2.21(2.27)	**1.74(1.96)**	1.86(1.82)	4.48(3.24)	0.69(1.33)
	985.00(0.00)	473.24(55.28)	985.00(0.00)	984.94(0.24)	822.50(4.77)
Lung Cancer	4.37(2.77)	5.49(3.18)	**4.22(2.66)**	8.69(3.80)	4.78(2.66)
	1000.00(0.00)	536.62(78.55)	1000.00(0.00)	999.78(0.42)	870.86(1.87)
Novartis	1.26(2.15)	**0.89(2.37)**	0.98(1.98)	3.81(4.43)	0.39(1.34)
	500.00(0.00)	238.46(32.60)	500.00(0.00)	499.96(0.20)	424.22(3.16)

Table 1. Results for gene expression data. For comparisson we also included LMNN. The top entry is the average test error followed by the STD in parentheses. Below the error rates the average number of non-zero weights, again followed by the STD, is given.

3 Experiments

Experiments on UCI datasets show that MDM is independent of the initial scaling of the data dimensions [6]. Here we applied it to gene expression datasets available from the Broad Institute webside[1]. The data dimensions of each dataset were normalized so that the data points have zero mean and a variance of one. The k-NN (k=3) error rates in Table 1 were obtained by a 5-fold cross-validation that was repeated ten times. None of the tested methods is clearly better then any other and the variances are quite large. This shows how challanging this data is. There are only 70 to 250 samples and it has 500 to 1200 dimensions. Interestingly, MDM reduces the dimensionality heavily, which is worth to have a closer look at.

References

1. Cover, T., Hart, P.: Nearest neighbor pattern classification. Information Theory, IEEE Transactions on **13**(1) (1967) 21–27
2. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. Neural Netw. **15**(8-9) (2002) 1059–1068
3. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proc. 9th International Workshop on Machine Learning. (1992) 249–256
4. Gilad-Bachrach, R., Navot, A., Tishby, N.: Margin based feature selection - theory and algorithms. In: Proceedings of the twenty-first international conference on Machine learning. ICML '04, New York, NY, USA, ACM (2004) 43–50
5. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press (2006)
6. Hocke, J., Martinetz, T.: Feature Weighting by Maximum Distance Minimization. In: ICANN 2013. (to appear)

[1] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi