Automatic Adjustment of Discriminant Adaptive Nearest Neighbor

Nicolas Delannay, Cédric Archambeau, Michel Verleysen*
Université catholique de Louvain (UCL)
DICE - Machine Learning Group
Place du levant 3, 1348 Louvain-la-Neuve, Belgium
Nicolas.Delannay@uclouvain.be, {archambeau,verleysen}@dice.ucl.ac.be

Abstract

K-Nearest Neighbors relies on the definition of a global metric. In contrast, Discriminant Adaptive Nearest Neighbor (DANN) computes a different metric at each query point based on a local Linear Discriminant Analysis. In this paper, we propose a technique to automatically adjust the hyper-parameters in DANN by the optimization of two quality criteria. The first one measures the quality of discrimination, while the second one maximizes the local class homogeneity. We use a Bayesian formulation to prevent overfitting.

1. Introduction

Consider a classification problem with J classes and N training observations. The observations have D features. The goal is to predict the class of a new observation at a query point x_0. K-Nearest Neighbors (K-NN) is a simple method to tackle classification. The principle is to find the K nearest points to x_0 in the training set and to predict that x_0 belongs to the most represented class within the nearest neighbor set. K-NN relies on the definition of a metric. The Euclidean metric assumes equal importance to all directions in space. This is a limiting assumption especially in high-dimensional space. A better metric should constrict distances in directions where class densities are constant and elongate distances in directions where class densities change. This idea lead in [4] to the development of the Discriminant Adaptive Nearest Neighbor (DANN). For a query point x_0, DANN finds anisotropic directions based on a locally weighted Linear Discriminant Analysis (LDA). These directions allow defining a local metric to be used in K-NN. A similar procedure is proposed in [3] with recursive partitioning methods. In [6] and [2], it is a support vector classifier that is used to construct the anisotropic metric. In [5], a soft version of K-NN is used with a kernel smoother.

While K-NN requires choosing only the number of neighbors, locally adaptive methods usually add several other hyper-parameters. These are set by heuristics or cross-validation techniques. In this paper, we follow DANN and propose a technique to optimize directly the hyper-parameters. This new method is called automatic DANN. The principle is to maximize two criteria with respect to the hyper-parameters. The first criterion evaluates the quality of discrimination of the local LDA model while the second one evaluates the homogeneity of the distribution of points in the final neighborhood. We will show that this distribution can be viewed as a local Multinomial model. To prevent overfitting, we work with a Bayesian formulation of the local models.

In section 2, DANN is exposed. In section 3, automatic DANN is developed and illustrated on a toy example. Section 4 presents a comparison of automatic DANN, LDA, K-NN and DANN on synthetic and real datasets.

2. Discriminant Adaptive Nearest Neighbors

DANN proceeds in two steps. In the first step, the class distributions around a query point x_0 are locally modeled with LDA. The second step computes a matrix Σ_{DANN} and uses the associated metric in K-NN.

Subsequently, a point from the training set is noted x_n and the corresponding class label is noted $t_n \in \{1, \ldots, J\}$. To define locality around query point x_0, DANN considers the K_Q nearest neighbors according to the Euclidean metric. The local LDA model computes then the within and between classes covariances by

$$\bar{W} = \frac{1}{N_Q} \sum_{j=1}^{J} \sum_{t_n=j} w_{Q_n} (x_n - \bar{\mu}_j)(x_n - \bar{\mu}_j)^{\mathrm{T}} \quad (1)$$

$$\bar{B} = \sum_{j=1}^{J} \pi_{Qj} (\bar{\mu}_j - \bar{\mu})(\bar{\mu}_j - \bar{\mu})^{\mathrm{T}} \ , \quad (2)$$

*N.D. and M.V. are respectively Research Fellow and Research Director of the Belgian National Fund for Scientific Research.

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

where the contribution of x_n to the local LDA is $w_{Qn} = 1$ if x_n belongs to the K_Q nearest neighbors of x_0 and $w_{Qn} = 0$ otherwise, $\bar{\mu}_j = \sum_{t_n=j} w_{Qn} x_n / N_{Qj}$ is the weighted mean of the observations in class j and $\bar{\mu}$ is the global weighted mean. $N_Q = \sum_n w_{Qn}$ is the equivalent number of points included in the local LDA model and $\pi_{Qj} = N_{Qj}/N_Q$ are the empirical prior class probabilities with $N_{Qj} = \sum_{t_n=j} w_{Qn}$.

Next, DANN considers the K_H nearest neighbors to x_0 and constructs a second K-NN classifier based on the weighted Euclidean metric

$$d_{\Sigma_{DANN}}(x, x_0)^2 = (x - x_0)^{\mathrm{T}} \Sigma_{DANN}^{-1} (x - x_0) \ , \quad (3)$$

where $\Sigma_{DANN}^{-1} \equiv \bar{W}^{-1} \bar{B} \bar{W}^{-1} + \epsilon \bar{W}^{-1}$. The hyper-parameter ϵ prevents the space to be totally shrunk in least discriminant directions (see [4] for further details).

DANN has thus three hyper-parameters: K_Q, ϵ and K_H. As such, the method does not give indications on how to set these parameters; we have to call upon heuristics or cross-validation techniques. Automatic DANN tries to overcome this limitation.

3. Automatic DANN

3.1. New parametrization

Instead of using hard neighborhoods, we prefer to work with soft ones. The resulting parametrization is more convenient when optimizing the hyper-parameters in automatic DANN.

To define local contributions w_{Qn}, we will use a gaussian weighting function instead of K_Q-NN,

$$w_{Qn} = \exp\Big(-\frac{1}{2} d_{\Sigma_Q}(x_n, x_0)^2\Big) \ . \quad (4)$$

In this paper, we work with $\Sigma_Q = \rho_Q I_D$ where I_D is the identity matrix and ρ_Q is a global scale parameter playing a role similar to K_Q.

Also, we do not work with Σ_{DANN} but a slightly different matrix Σ_H. Let us note U_Q the matrix of eigenvectors of $\bar{W}^{-1}\bar{B}$ corresponding to the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_D$. According to LDA, the eigenvectors in U_Q are the successive most discriminant directions. For this reason, we chose to express Σ_H directly with the LDA eigen values/vectors decomposition:

$$\Sigma_H = \rho_H U_Q \, \Delta^\gamma \, U_Q^{\mathrm{T}} \ , \quad (5)$$

where Δ is a diagonal matrix with diagonal elements δ_d inversely proportional to the eigenvalues ($\delta_d \propto 1/\lambda_{Qd}$) and with additional constraint $\Pi_d \delta_d = 1$. We impose a priori a bound on the anisotropy to prevent some directions to be completely degenerated:

$$\forall d, \ \lambda_{Q1} > \lambda_{Qd} \cdot R_{max} : \ \delta_d = \delta_1 \cdot R_{max} \ . \quad (6)$$

In (5), matrix Σ_H is parameterized by two new hyper-parameters, γ and ρ_H, playing roles similar to ϵ and K_H.

Finally, it is convenient to view the K_H-NN classifier as a local Multinomial model. The contributions w_{Hn} of the training points to this model are computed by (4) with the distance $d_{\Sigma_H}(x_n, x_0)$. The number $N_{Hj} = \sum_{t_n=j} w_{Hn}$ of local points in each class are drawn from a Multinomial distribution :

$$p(\{N_{Hj}\}) = \mathcal{M}n(N_H, \{\pi_{Hj}\}) \ , \quad (7)$$

where $N_H = \sum_n w_{Hn}$ refers to an equivalent number of points in the local Multinomial model and the π_{Hj} are the unknown prior class probabilities. In DANN (or K-NN in general), the prior class probabilities are evaluated by the empirical means $\pi_{Hj} = N_{Hj}/N_H$ and the predicted class at x_0 is given by $y(x_0) = \mathrm{argmax}_j \pi_{Hj}$.

In the next section, we show how to optimize the hyper-parameters ρ_Q, γ and ρ_H.

3.2. Optimization Criteria

In automatic DANN, we propose to adjust the hyper-parameters in two stages. First, we optimize ρ_Q based on the quality of discrimination of the local LDA model. This model gives us an orientation U_Q and relative scale Δ to construct Σ_H. To complete the definition, we set γ and ρ_H in order to maximize the homogeneity according to the local Multinomial model.

The quality of discrimination can be measured by a weighted classification likelihood

$$Q(\rho_Q) = \frac{1}{J} \sum_{j=1}^{J} \sum_{t_n=j} \frac{w_{Qn} \, p(y_n = j | x_0, \mathcal{M}_Q)}{N_{Qj}} \ , \quad (8)$$

where $\mathcal{M}_Q$ refers to the local LDA. Following this model, we have $p(y_n = j | x_0, \mathcal{M}_Q) \propto \mathcal{N}(x_n | \mu_j, \bar{W}) \ \pi_{Qj}$, with $\mathcal{N}(\cdot|\cdot,\cdot)$ the multivariate normal distribution.

The weighting appearing in (8) accounts for local contribution of points and gives equal importance to each class. This second property sets the quality of discrimination to $1/J$ (i.e., random guess) when only one class is represented locally.

The homogeneity of the local Multinomial model is defined by

$$H(\rho_H, \gamma) = \max_j (\pi_{Hj}) \ . \quad (9)$$

The maximization of the homogeneity corresponds to finding a neighborhood around x_0 in which one class is highly dominant.

Working with too few points in the local models leads to overfitting. For example, a local LDA computed with one point in each class has a high $Q(\cdot)$ although the discriminant directions U_Q are not relevant for the complete distribution. Also, at the limit $\rho_H \rightarrow 0$, the neighborhood corresponds to 1-NN for which homogeneity is perfect. To circumvent this problem, we introduce a Bayesian formulation of the local models.

3.3. Bayesian formulation

Following the Bayesian paradigm, we will assume that the distribution of the parameters of the local models are conjugate prior distributions. This leads to tractable posterior distributions. We will not work with the complete posterior distributions as marginalisation integrals are too expensive. Instead, we consider the posterior means and use these estimates of the model parameters to evaluate the criteria (8) and (9).

We will assume that the parameters of the LDA model $\{W, \mu_j, \pi_{Qj}\}$ are random variables with prior Normal-Wishart and Dirichlet distributions

$$p(\{\mu_j\}, W^{-1}) = \mathcal{W}i(W^{-1}|\nu^0, W^0) \prod_j \mathcal{N}(\mu_j|\mu_j^0, N_{Qj}^0 W)$$
$$p(\{\pi_{Qj}\}) = \mathcal{D}i([N_{Q1}^0, \ldots, N_{QJ}^0]) \ . \tag{10}$$

Then, we can show (e.g. [1]) that the posterior distributions of the parameters have a known analytical forms, and the corresponding mean posterior estimates are

$$\mu_j^\star = (N_{Qj}^0 \mu_j^0 + N_{Qj}\bar{\mu}_j)/(N_{Qj}^0 + N_{Qj})$$
$$W^\star = \frac{W^0 + \frac{1}{2}\bar{W} + \frac{1}{2}\sum_j \frac{N_{Qj}^0 N_{Qj}}{N_{Qj}^0 + N_{Qj}}(\mu_j^0 - \bar{\mu}_j)(\mu_j^0 - \bar{\mu}_j)^{\mathrm{T}}}{\nu^0 + \frac{1}{2}(N_Q - D - 1)}$$
$$\pi_{Qj}^\star = (N_{Qj}^0 + N_{Qj})/(N_Q^0 + N_Q) \ . \tag{11}$$

The choice of the parameters of the prior reflects additional model assumptions. Considering that a priori we only know the position of the query point x_0 and the metric defined by Σ_Q, we propose to set

$$W^0 = \Sigma_Q, \quad \nu^0 = \frac{1}{2}(D+1) + J^2,$$
$$\mu_j^0 = x_0, \quad N_{Qj}^0 = N_Q^0/J \ . \tag{12}$$

The two first equalities set the prior mean of W to Σ_Q/J^2 and thus the prior volume for each class is $\sqrt{\rho_Q}/J$. With the third equality we assume the distributions to be centered on the query point. The last equality gives equal importances to all the classes. The parameter N_Q^0 corresponds to an equivalent number of points attributed to the prior. Additional knowledge should help choosing this parameter.

The hyper-parameter ρ_Q is selected by (8) and we compute the eigenvectors U_Q and eigenvalues λ_d of $W^{\star -1}B^\star$, where $B^\star$ comes from (2) with $\mu_j^\star$ and $\pi_{Qj}^\star$.

For the Multinomial model, we impose a Dirchlet distribution on the class probabilities $p(\{\pi_{Hj}\}) = \mathcal{D}i(\{N_{Hj}^0\})$. It is natural to impose $N_{Hj}^0 = N_H^0/J$ where N_H^0 corresponds to an equivalent number of points attributed to the prior. The posterior distribution become a Dirichlet, with mean

$$\pi_{Hj}^\star = (N_{Hj}^0 + N_{Hj})/(N_H^0 + N_H) \ . \tag{13}$$

The hyper-parameters γ and ρ_H are optimized with (9) and the predicted class at x_0 is $y(x_0) = \mathrm{argmax}_j \pi_{Hj}^\star$.

We can see in (11) and (13) that the posterior mean estimates corresponds to the empirical estimates used in DANN, biased towards the means of the prior distributions.

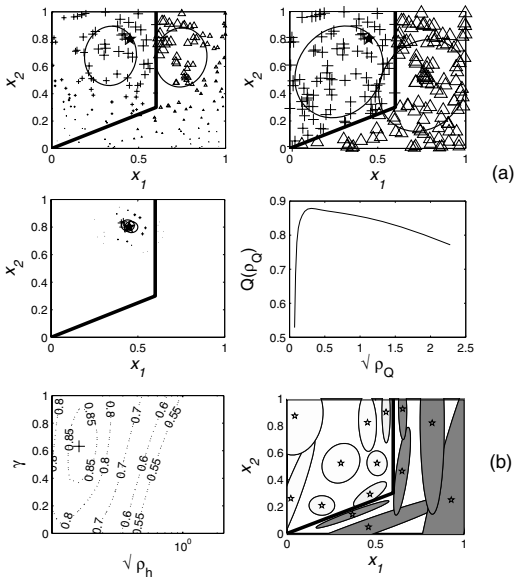

Figure 1. (a) Local LDA model for $\sqrt{\rho_Q} = \{0.15, 0.5, 1.5\}$ and corresponding $Q(\rho_Q)$. (b) (left) Homogeneity criterion $H(\gamma, \rho_H)$ and (right) shape of final neighborhoods.

3.4. Illustration

Automatic DANN is illustrated on a two-dimensional toy example. On figure 1 (a) we can visualize the local LDA model for three values of ρ_Q. The two segments line is the true classification boundary. The star represents the query point and the ellipses represent the estimated class distributions (centered in $\mu_j^\star$ and with the shape determined by $W^\star$). The size of the markers accounts for the contributions w_{Qn} of the training points to the local model. The bottom right plot shows the corresponding criterion $Q(\rho_Q)$.

On figure 1 (b) left, $H(\gamma, \rho_H)$ is evaluated for the same query point on a grid of hyper-parameter values. On the right plot, we visualize the shape of the neighborhoods defined by Σ_H for different query points.

4. Experiments and Discussion

In this section, we compare LDA, K-NN, DANN and our automatic DANN method (a-DANN) on four datasets discribed below. For LDA, the class priors is determined by the empirical estimates. For K-NN, the Euclidean distance is used and the number K is optimized by 5-fold cross-validation. For DANN, we follow [4] and fix the hyper-parameters K_Q to $\min\{50, N/5\}$ and ϵ to 1. The hyper-parameter K_H is optimized by 5-fold cross-validation. For a-DANN, the anisotropy is limited to $R_{max} = 16$. We also took $N_Q^0 = K_Q/3$ and $N_H^0 = K/3$. This choice affects the optimization of the hyper-parameters. Only further model assessement can tell us the extend of this influence. However, it should be small if the data distribution matches the model assumptions.

4.1. Datasets

- *Two normals*: two normaly distributed classes in dimension two centered at (0,0) and (2,0). Predictors have variance (1,2) and correlation 0.75. There are 14 additional standard Gaussian noise features.
- *Radial*: four-dimensional data with the same number of data in each class. The first class is generated by $\mathcal{N}(0, I)$. The second class is isotropic around the origin 0 and the norm of the data are drawn from an uniform distribution $\mathcal{U}(2.6, 3.5)$.

For these simulated data, we generate 20 times 200 points for learning and 300 points for test.

- *Sonar*: 60 features reduced to 10 by PCA, two classes to predict (*mine* or *rocks*) and 208 observations.
- *Glass*: 9 features to predict, 6 classes (the type of glass) and 214 observations.

For these two datasets taken from the UCI repository, we repeat 20 times a random 5-fold cross-validation and compute for each run the average validation error.

For all the experiments, the data are normalized before running the algorithms. The 50 percent central quantiles on misclassification rates are reported in table 1.

4.2. Discussion

The *two normals* dataset respects the assumptions of the LDA model. As expected, LDA gives the best performances. What is interesting is that a-DANN is approaching the LDA performances while DANN is not.

On the *radial* dataset, LDA is incapable of discriminating the classes. On the contrary, the local LDA can bring useful information as the DANN and a-DANN methods perform better than K-NN.

In all the experiments, automatic DANN performs consistently better than DANN. However, it is seen in the *Sonar* and *Glass* results that the advantage of using DANN and a-DANN is not obvious on real datasets in such dimensionality. We have the feeling that the method would benefit from additional constraints reducing the effective space dimensionality.

	LDA	K-NN	DANN	a-DANN
2 Normals	7-8.5	19-23	17-21	9.8-12
Radial	47-51	20-23	15-18	13-16
Sonar	20-23	14-16	16-18	14-16
Glass	37-39	29-32	33-36	30-33

Table 1. Percentage of misclassifications

5. Conclusion

We propose a method to automatically adjust the hyper-parameters of Discriminant Adaptive Nearest Neighbor. The method is based on the definition of two criteria for evaluating the quality of local LDA and local Multinomial models. To prevent overfitting appearing with local models, we introduce prior distributions over the parameters and work with the posterior mean estimates.

We show on simple classification examples the advantage of adjusting the hyper-parameters over an heuristic. Further research should look for additional constraints that effectively reduce space dimensionality.

References

[1] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1994.

[2] C. Domeniconi, D. Gunopulos, and J. Peng. Large margin nearest neighbor classifiers. *IEEE Transactions on Neural Networks*, 16(4):899–909, July 2005.

[3] J. Friedman. Flexible metric nearest neighbor classification. Technical report, 1994.

[4] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE transaction on Pattern Analysis and Machine Intelligence*, 18(6):607–615, 1996.

[5] D. G. Lowe. Similarity metric learning for a variable-kernel classifier. *Neural Computation*, 7(1):72–85, January 1995.

[6] J. Peng, D. R. Heisterkamp, and H. K. Dai. Lda/svm driven nearest neighbor classification. In K. Marriott, editor, *IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.