Annealed Discriminant Analysis

Gang Wang, Zhihua Zhang, and Frederick H. Lochovsky

Department of Computer Science,
Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong
{wanggang, zhzhang, fred}@cs.ust.hk

Abstract. Motivated by the analogies to statistical physics, the deterministic annealing (DA) method has successfully been demonstrated in a variety of applications. In this paper, we explore a new methodology to devise the classifier under the DA method. The differential cost function is derived subject to a constraint on the randomness of the solution, which is governed by the temperature T. While gradually lowering the temperature, we can always find a good solution which can both solve the overfitting problem and avoid poor local optima. Our approach is called *annealed discriminant analysis* (ADA). It is a general approach, where we elaborate two classifiers, i.e., distance-based and inner product-based, in this paper. The distance-based classifier is an annealed version of linear discriminant analysis (LDA) while the inner product-based classifier is a generalization of penalized logistic regression (PLR). As such, ADA provides new insights into the workings of these two classification algorithms. The experimental results show substantial performance gains over standard learning methods.

1 Introduction

The deterministic annealing (DA) technique has demonstrated substantial performance improvement over clustering, classification and constrained optimization problems [1, 2, 3, 4, 5]. Since DA is strongly motivated by the analogies to statistical physics [6], it regards the optimization problem in question as a thermal system. The Lagrange multiplier in the problem represents the temperature of the system, which is used to control the level of randomness, and the cost function corresponds to the free energy of the system. The minimum of the free energy determines the state of the system at thermal equilibrium. To achieve the equilibrium state, one tracks the minimum of the free energy while gradually lowering the temperature. At the limit of low temperature, minimum energy is reached. In other words, the DA technique performs annealing as it maintains the cost function at its minimum while gradually lowering the temperature. With careful annealing, this process can avoid many shallow local minima of the specified cost and finally produce a non-random solution. The DA technique is attractive since it possesses two important advantages: (1) the ability to minimize the cost function even when its gradients vanish almost everywhere; (2) the ability to avoid many poor local optima.

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, pp. 449–460, 2005.

Since direct classification error minimization mostly leads to an NP-hard problem [7], the goal of the learning methods is to avoid the computational difficulties of this hard problem. Usually, we transform the learning problem into an optimization problem, where an objective function is proposed. With different criteria, such as maximum likelihood, maximum posterior estimation, least L-norm error, or maximum margin, we can construct different classifiers. Wabha [8] treated these classifiers as performing "soft" or "hard" classifications. Soft classification, such as logistic regression models, assigns an object based on a conditional probability of this object in some class, while hard classification, such as SVMs, does not use the probability. We bridge the gap between these two kinds of classification problems through the DA approach. Instead of treating the "soft" and "hard" separately, we formulate the objective by a "hard" notion, and solve it by a "soft" way.

In this paper, we formulate the classification problem based on the discriminant functions. The optimal hypotheses is hard to find directly, since the problem is both a "hard" classification problem and NP-hard. However, with the introduction of a conditional probability in DA, the classification becomes soft. Hence, the original non-differentiable cost function that results in an NP-hard problem becomes differentiable. Interestingly, this "soft" problem also tends to the original "hard" problem as the temperature approaches zero. In addition, Rose [4] argued that the entropy can play a role in the regularization. Therefore, motivated by these observations, we investigate the application of the DA technique to the classification problem and devise a new kind of method called annealed discriminant analysis (ADA). Since ADA is a general formulation, we present two possible implementations, i.e., distanced-based classifier and inner product-based classifier. The distanced-based classifier is closely related to linear discriminant analysis (LDA), since the parameters of LDA are real means of the categories while the parameters of the distanced-based classifier are the "soft" means, which are estimated from the iterative updating procedure. Thus, the distanced-based classifier can be seen as an annealed version of LDA. The inner product-based classifier is a generalization of penalized logistic regression (PLR) since they become the same when setting the temperature to one. Therefore, ADA provides new insights into the workings of these existing classification algorithms.

The rest of this paper is organized as follows. Section 2 derives the ADA algorithm based on the discriminant functions and the DA approach. Two implementations of ADA are elaborated in section 3. Section 4 reports our experiment setup and results. The last section presents the concluding remarks.

2 Problem Formulation

Let $\mathcal{T} = \{(\mathbf{x}_i, c_i)\}$ be a training set of N labelled vectors, where $\mathbf{x}_i \in \mathbf{R}^d$ is a feature vector and $c_i \in \mathcal{I}$ is its class label from an index set $\mathcal{I} = \{1, 2, \ldots, C\}$. A classifier is a mapping $F : \mathbf{R}^d \to \mathcal{I}$, which assigns a class label in $\mathcal{I}$ to each vector in $\mathbf{R}^d$. A training pair $(\mathbf{x}, c) \in \mathcal{T}$ is correctly classified if $F(\mathbf{x}) = c$.

We code c_i as a binary C-vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iC})'$ with values all zero except a 1 in position c if the class is c. We shall interchangeably use c_i and $\mathbf{y}_i$ to indicate the class label of $\mathbf{x}_i$ in this paper.

2.1 Problem Definition

We formulate a classifier in terms of a set of discriminant functions $\{g(\mathbf{x};\boldsymbol{\theta}_j) \mid j = 1, 2, \ldots, C\}$ such that an input vector $\mathbf{x}$ is assigned to the class c if and only if

$$g(\mathbf{x};\boldsymbol{\theta}_c) \geq g(\mathbf{x};\boldsymbol{\theta}_j) \text{ for all } j \neq c, \tag{1}$$

where the parameter $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j\}_{j=i}^{C}$ is a set of vectors for indexing the discriminant functions. Here the discriminant functions are general, which can be directly defined as many function forms such as Gaussian, linear, etc. The above classification rule defines a "hard" classification [8]. Denoting the conditional probability of the class c given $\mathbf{x}_i$ by $p_{ic} = p(c|\mathbf{x}_i)$, this "hard" classification implies that

$$p_{ic} = \begin{cases} 1 \; c = \mathrm{argmax}_j g(\mathbf{x}_i;\boldsymbol{\theta}_j) \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

Let $\mathbf{g}_i = (g(\mathbf{x}_i;\theta_1), g(\mathbf{x}_i;\theta_2), \ldots, g(\mathbf{x}_i;\theta_C))'$ and $\|\mathbf{g}_i\|_\infty = \max\{g(\mathbf{x}_i;\boldsymbol{\theta}_j)\}_{j=1}^{C}$. For a hypotheses indexed by the parameter $\hat{\boldsymbol{\theta}}$, if a case $\mathbf{x}_i$ has correctly been classified according to (1), we have $\|\mathbf{g}_i\|_\infty = \mathbf{y}_i'\mathbf{g}_i$. Otherwise we have $\|\mathbf{g}_i\|_\infty > \mathbf{y}_i'\mathbf{g}_i$. This leads us to define a classification error

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_i \Big|\|\mathbf{g}_i\|_\infty - \mathbf{y}_i'\mathbf{g}_i\Big| = \frac{1}{N}\sum_i \Big(\|\mathbf{g}_i\|_\infty - \mathbf{y}_i'\mathbf{g}_i\Big). \tag{3}$$

If all training samples have been correctly classified by this classifier, (3) will arrive at its minimum zero. Hence our task is to find a set of parameters $\{\boldsymbol{\theta}_j\}$ minimizing the classification error. Alternatively, with the above definition for p_{ic}, we can rewrite (3) as

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_i \big(\sum_j p_{ij} g(\mathbf{x}_i;\boldsymbol{\theta}_j) - \mathbf{y}_i'\mathbf{g}_i\big) \tag{4}$$

due to $\sum_j p_{ij} g(\mathbf{x}_i;\boldsymbol{\theta}_j) = \|\mathbf{g}_i\|_\infty$. Minimizing the classification error (3) or (4) w.r.t. $\boldsymbol{\theta}$ requires searching all possible "hard" conditional probabilities, and therefore results in an NP-hard problem. To find an approximate searching strategy to obtain the best parameter $\boldsymbol{\theta}$ to minimize the classification error (4) is also not straightforward because this error function is non-differentiable. Even if we get a solution, it often suffers from the overfitting problem. Thus, our current problem is how to search in the parameter space to get a good solution. The essence of the DA technique [4] is to cast the optimization problem into a probabilistic framework, considering a "randomness" characterized by a probabilistic assignment of data to classes. The DA approach is a good choice to deal with these problems.

2.2 Deterministic Annealing Approach

Recall that the minimization of $L(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ is intractable since $L(\boldsymbol{\theta})$ is non-differentiable. Our departure is replacing the discrete $\{p_{ij}\}$ with a continuous density function $\{p_{ij}\}$[1] and finding a differential function to approximate $L(\boldsymbol{\theta})$. Since $\{p_{ij}\}$ are now unknown continuous density distribution functions, our problem is how to select $\{p_{ij}\}$. Let

$$E = \sum_i \sum_j p_{ij} g(\mathbf{x}_i; \boldsymbol{\theta}_j) \tag{5}$$

E is a variational function where its parameters $\{p_{ij}\}$ are determined by the discriminant functions. From the definition of the conditional probability in (2), $\{p_{ij}\}$ need to maximize E given a hypotheses. Therefore we are seeking $\{p_{ij}\}$ maximizing E subject to a specified level of randomness measured by the Shannon entropy while assuming the parameters of the discriminant functions are fixed. The entropy is defined as

$$H = -\sum_i \sum_j p_{ij} \log p_{ij}.$$

Maximum entropy is inspired by the well known principle of Occam's razor, which states that the simplest model that accurately represents the data is the most desirable. This criteria tends to induce the parsimony model to fit the data. Conveniently, this optimization is reformulated as maximization of the Lagrangian

$$F = E + TH \tag{6}$$

where T is the Lagrange multiplier. For large value of T, the probabilities mainly attempt to maximize the entropy, and as T approaches zero, it maximizes E.

Maximizing F w.r.t. p_{ij} is straightforward, giving rise to the Gibbs distribution [9]

$$p_{ij} = \frac{\exp(\frac{g(\mathbf{x}_i;\boldsymbol{\theta}_j)}{T})}{\sum_k \exp(\frac{g(\mathbf{x}_i;\boldsymbol{\theta}_k)}{T})}. \tag{7}$$

The corresponding maximum of F is obtained by plugging (7) back into (6)

$$F^* = \max_{\{p_{ij}\}} F = T \sum_i \log \sum_j \exp(\frac{g(\mathbf{x}_i; \boldsymbol{\theta}_j)}{T}). \tag{8}$$

It is easy to see

$$T \log \sum_j \exp(\frac{g(\mathbf{x}_i; \boldsymbol{\theta}_j)}{T}) \geq T \log \exp(\frac{\mathbf{y}_i' \mathbf{g}_i}{T}) = \mathbf{y}_i' \mathbf{g}_i.$$

[1] The "hard" conditional probability p_{ij} takes binary values. For notation simplicity, we still denote the soft conditional probability as p_{ij}.

Hence replacing $\|\mathbf{g}_i\|_\infty$ with $T \log \sum_j \exp(\frac{g(\mathbf{x}_i;\boldsymbol{\theta}_j)}{T})$ in $L(\boldsymbol{\theta})$, we obtain a differential cost function:

$$Q(\boldsymbol{\theta}) = \frac{1}{N}\sum_i \left[T \log \sum_j \exp(\frac{g(\mathbf{x}_i;\boldsymbol{\theta}_j)}{T}) - \mathbf{y}_i'\mathbf{g}_i\right] \tag{9}$$

to approximate $L(\boldsymbol{\theta})$. We address the optimization problem on minimization of $Q(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ alternatively.

Theorem 1. *Let p_{ij} and $Q(\boldsymbol{\theta})$ be defined by (7) and (9), respectively. For a fixed parameter $\boldsymbol{\theta}$,*

1. $\lim_{T\to\infty} p_{ij} = \frac{1}{C}$ *for* $j = 1, \ldots, C$.
2. $\lim_{T\to 0} p_{ic} = \begin{cases} 1 & c = argmax_j g(\mathbf{x};\boldsymbol{\theta}_j) \\ 0 & otherwise; \end{cases}$,
   $\lim_{T\to 0} Q(\boldsymbol{\theta}) = L(\boldsymbol{\theta})$;
3. $Q(\boldsymbol{\theta})$ *is monotone decreasing with respect to decreasing* T.

We omit the proof due to the space limitations. This theorem states that at infinite temperature T, the conditional probabilities $\{p_{ij}\}$ are soft as they are uniformly distributed for all categories. At the limit of zero temperature, the classification is hard where each case is assigned to the category whose discriminant function value is largest. When the conditional probabilities $\{p_{ij}\}$ become hard, F in (6) tends to E in (5), and consequently the cost function $Q(\boldsymbol{\theta})$ in (9) degenerates to $L(\boldsymbol{\theta})$ in (3). Thus, the original problem is in fact a "zero temperature problem". This motivates a criterion for updating T: start with a high value of temperature T and track the minimum while lowering T. Since $Q(\boldsymbol{\theta})$ is monotone decreasing with respect to the temperature T, the above algorithm will try to converge to a global minimum. The entire algorithm is presented in Table 1.

Table 1. A brief sketch of the ADA algorithm

```
0. Initialize $T$ with a comparatively large value $T^{(0)}$
1. Initialize $\boldsymbol{\theta}^{(0)}$
2. Repeat
3.     Lower temperature: $T^{(t)} = q(T^{(t-1)})$
4.     $\boldsymbol{\theta}^{(t)} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^{(t-1)})$
5.     Validate the performance for $(\boldsymbol{\theta}^{(t)}, T^{(t)})$
6. Until the parameters converge
7. Select $(\boldsymbol{\theta}^{(t)}, T^{(t)})$
8. Classify cases based on (7)
```

The algorithm consists of two-level iterations. In the inner iteration,for a fixed T, we optimize the $Q(\boldsymbol{\theta})$. We can not get the closed form to update the parameters, hence we resort to a numerical optimizer, such as the conjugate gradient algorithm, to find the parameter values.

For the outer iteration, in our experiments, we use an exponential schedule for reducing T, i.e., $q(T) = \alpha T$, where $\alpha < 1$ in our experiments. From the perspective of the learning problem, we start with a very simple model with zero variance. We then gradually increase the complexity of the model. Since the bias would reduce faster than the variance increases, thus the prediction error would decrease also. However, when the variance increases faster than the bias at a certain temperature, overfitting occurs. Therefore, the final converging parameters from "hard“ partitions may not reach the best performance due to the overfitting problem. Consequently, we need to select parameters according to their performance on the validation set. Empirically, the optimal parameter can be obtained at the temperature that causes the conditional probabilities to be almost "hard" while still a little "soft". Since the temperature T controls the level of the randomness of the solution, we try the model to fit the data with different complexity from simple to complex as the temperature decreases. Therefore, we are certain to find an optimal classifier based on the ADA algorithm.

[2, 4] also discuss the supervised learning problem using DA methods, but they are different from our ADA method. In their work, they construct the space partitioning functions as two components: a parametric space partition (structured partition) and a parametric local model per partition cell, while in our approach we integrate them by using only the discriminant function. Furthermore, they employ the notion of the regression problem by defining the distortion to derive the learning algorithm, while ADA directly formulates the classification problem by a discriminant function.

3 Annealed Discriminant Analysis

In this section, we will discuss two implementations in ADA, i.e., distance-based classifier and inner product-based classifier. The first one is related to linear discriminant analysis(LDA) and the second is related to penalized logistic regression(PLR).

3.1 Distance-Based Classifier (dADA)

The discriminant functions are defined as

$$g(\mathbf{x}; \boldsymbol{\mu}_j) = -(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \; j = 1, \ldots, C \tag{10}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix which is evaluated in advance from the training data, and the parameter $\boldsymbol{\mu}_j \in \mathbf{R}^d$ indicates the mean of the j-th category. A case $\mathbf{x}$ will be classified to the category c_j when the distance between its mean $\boldsymbol{\mu}_j$ and $\mathbf{x}$ is the least, and correspondingly the j-th discriminant function g_j is the largest. The decision boundary, corresponding to $g(\mathbf{x}; \boldsymbol{\mu}_k) = g(\mathbf{x}; \boldsymbol{\mu}_j)$, forms a hyperplane, i.e.,

$$f(\mathbf{x}) = \mathbf{x} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) + \boldsymbol{\mu}_k' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j = 0 \tag{11}$$

If $f(\mathbf{x}) > 0$, the case $\mathbf{x}$ will be classified to the j-th category. Otherwise, $\mathbf{x}$ will be classified to the k-th category.

To find the optimal parameter $\boldsymbol{\mu}^{(t)}$ such that $\boldsymbol{\mu}^{(t)} = \arg\min_{\mu} Q(\boldsymbol{\mu})$ at a certain temperature, the gradient of the cost function Q is

$$\frac{\partial Q}{\partial \boldsymbol{\mu}_j} = \frac{2}{N} \sum_i [(p_{ij} - y_{ij}) \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j)] \tag{12}$$

where p_{ij}, with the form in (7), contains the parameter $\boldsymbol{\mu}$ via the discriminant functions. Since we cannot find the closed form for updating the parameters, we use the scaled conjugate gradient (SCG) optimizer [10], which is an extremely efficient algorithm, to get the parameters.

In LDA [11], cases in each category are assumed from a multivariate Gaussian, and all Gaussian distributions have a common covariance matrix $\boldsymbol{\Sigma}$. So the discriminant functions of LDA are defined the same as in (10) where $\boldsymbol{\mu}_j$ is estimated as the mean of the j-th category and the covariance $\boldsymbol{\Sigma}$ is evaluated from the training data. We note that the only difference between dADA and LDA is in how to estimate the parameters. The parameters in LDA are the real means of categories, while the means of dADA are the "soft" means, which are estimated from the iterative updating procedure. Therefore, the distance-based classifier can be seen as an annealed version of LDA. The experiment result in the next section shows that dADA performs much better than LDA.

3.2 Inner Product-Based Classifier (pADA)

The discriminant functions are defined as

$$g(\mathbf{x}; \boldsymbol{\lambda}_j) = \boldsymbol{\lambda}_j' \mathbf{x} \; j = 1, \ldots, C \tag{13}$$

where $\boldsymbol{\lambda}_j \in \mathbf{R}^d$ is the parameter of the inner product-based classifier. In this case, the decision boundary is also a hyperplane. Plugging (13) back into the cost function Q (9) we get

$$Q(\boldsymbol{\theta}) = \frac{T}{N} \sum_i [\log \sum_j \exp(\frac{\boldsymbol{\lambda}_j' \mathbf{x}_i}{T}) - \sum_j y_{ij} \frac{\boldsymbol{\lambda}_j' \mathbf{x}_i}{T}] \tag{14}$$

The gradient of the cost function Q is

$$\frac{\partial Q}{\partial \boldsymbol{\lambda}_j} = \frac{1}{N} \sum_i [(p_{ij} - y_{ij}) \mathbf{x}_i] \tag{15}$$

The parameter estimation in this problem is different from the distance-based classifier. From the gradient in (15), we notice that the conditional probabilities p_{ij} are constrained to be equivalent to the supervised label y_{ij} when minimizing the cost function Q. Since the discriminant functions $\{g(\mathbf{x}; \boldsymbol{\lambda}_j)\}$ can take any value ranging from negative infinity to positive infinity, the temperature in this definition can not control the randomness of the conditional probabilities. When the temperature is extremely high, the parameter $\boldsymbol{\lambda}$ will also take a large value to make $p_{ij} = y_{ij}$. Therefore, annealing has no effect no matter what is the temperature. The inner product discriminant functions seem to absorb the temperature T into the parameters $\boldsymbol{\lambda}$, so the cost function in (14) can be simplified as

$$Q(\boldsymbol{\theta}) = \frac{1}{N}\sum_i[\log\sum_j \exp(\boldsymbol{\lambda}_j'\mathbf{x}_i) - \sum_j y_{ij}(\boldsymbol{\lambda}_j'\mathbf{x}_i)] \tag{16}$$

This cost function is exactly the same as the cost function of logistic regression [12], which has been widely studied in statistics. However, since the optimal parameters of this cost function (16) try to satisfy $p_{ij} = y_{ij}$ directly, this formulation suffers from the overfitting problem.

To overcome this problem, we need to limit the range of the discriminant functions. We add a penalty term to the cost function such that

$$Q_{new}(\boldsymbol{\theta}) = \frac{1}{N}\sum_i[T\log\sum_j \exp(\frac{\boldsymbol{\lambda}_j'\mathbf{x}_i}{T}) - \sum_j y_{ij}(\boldsymbol{\lambda}_j'\mathbf{x}_i)] + \frac{\varepsilon}{2}\sum_j \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j \tag{17}$$

where the regularization parameter ε controls the range of the values of the parameters $\boldsymbol{\lambda}$. Consequently, the temperature T again governs the level of the randomness of the solution. The gradient of the cost function Q_{new} (17) becomes

$$\frac{\partial Q_{new}}{\partial \boldsymbol{\lambda}_j} = \frac{1}{N}\sum_i[(p_{ij} - y_{ij})\mathbf{x}_i] + \varepsilon\sum_j \boldsymbol{\lambda}_j \tag{18}$$

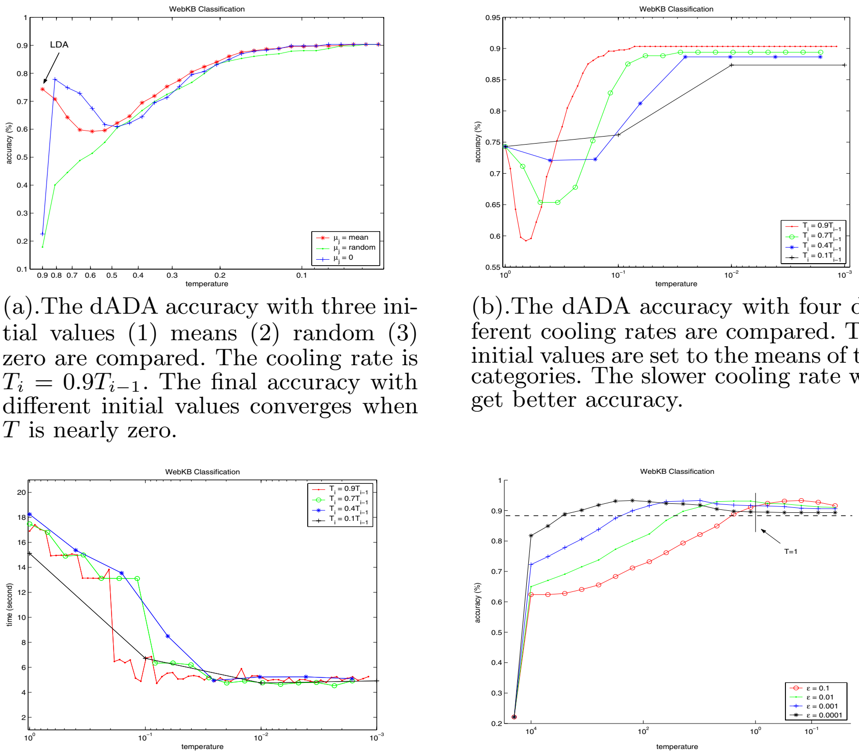
We also use the SCG optimizer to search for the optimal parameters in this classifier.

Penalized logistic regression (PLR) [13] began to gain attention recently because it not only performs as well as the SVM in two class classification, but can also naturally be generalized to the multi-class case. Furthermore, PLR provides an estimate of the conditional probability. As we can see, the cost function in (17) is the same as the negative log-likelihood of PLR when setting the temperature $T = 1$. Therefore, PLR is a special case of the pADA. Our approach gives a clear physical interpretation for PLR, where the temperature T controls the randomness of the solution, and the regularization parameter ε limits the range of the parameters. It is always laborious to select a good regularization parameter in PLR. We will see in the experiment that our algorithm can always find the optimal solution regardless of the value of the the regularization parameter, thereby, avoiding many attempts to examine it.

4 Experimental Results

4.1 WebKB: Web Pages Collection

The WebKB data set is a medium size collection, containing web pages gathered from several universities' computer science departments. The pages are divided into seven categories: student, faculty, staff, course, project, department and other. In this paper, we use the four most populous entity-representing categories: student, faculty, course, and project, which all together contain 4199 pages. A held-out set with 20% of the data was selected randomly. The other 80% was used as training data. Reserving those terms occurring at least six times in a corpus, we have 3359 training documents with a vocabulary size of 7161. We then use information gain to select 500 of the most predictive features and delete the cases without features.



(a).The dADA accuracy with three initial values (1) means (2) random (3) zero are compared. The cooling rate is $T_i = 0.9T_{i-1}$. The final accuracy with different initial values converges when T is nearly zero.

(b).The dADA accuracy with four different cooling rates are compared. The initial values are set to the means of the categories. The slower cooling rate will get better accuracy.

(c).The dADA time complexity with different cooling rates, where the setting is the same as in (b). The cost decreases as temperature is low.

(d).The pADA accuracy based on different regularization parameters. The dashed line is the accuracy of the logistic regression without regularization.

Fig. 1. The experiments on the WebKB dataset. The x-axis is a logarithm scale.

In the distance-based classifier, we use three ways to initialize the parameter values: (1) mean of each category; (2) a random number from all possible feature values; (3) zero. From Figure 1(a), we can see although these three lines have with different start points and different convergence traces, they merge in the end. The final accuracy of the ADA approach in fact is independent of the initial values. However, this is true only when we use a slow temperature cooling rate.

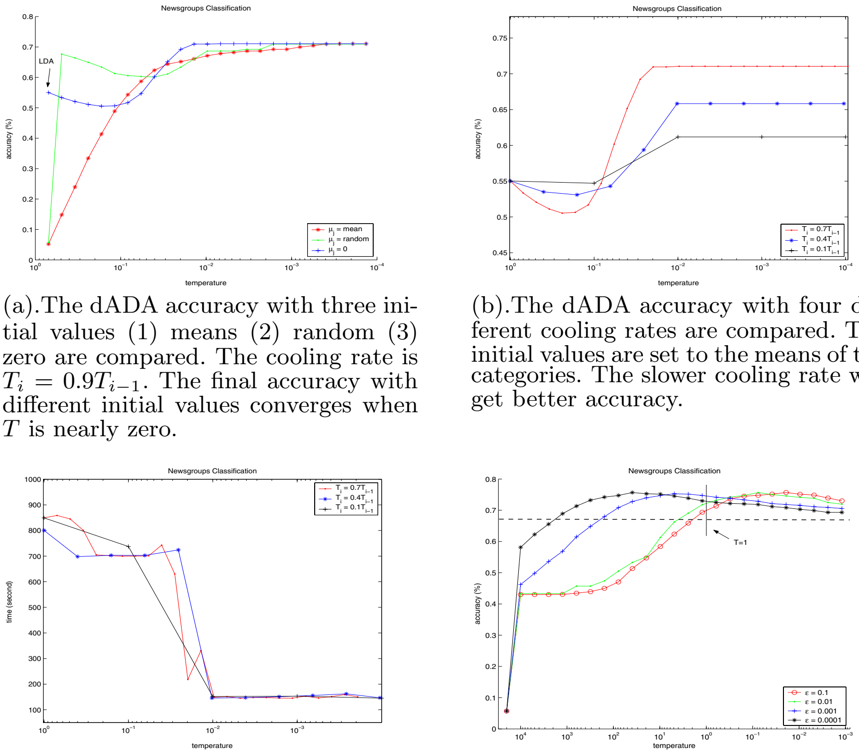
In this experiment, we lower the temperature with a comparatively slow rate, such as $T_i = 0.9T_{i-1}$. When the parameters are initialized to the means of the categories, the classifier becomes LDA, and its classification accuracy is 74.3%, much lower than the accuracy (90.3%) obtained by dADA. The LDA line is concave shaped. It goes down in the beginning since high temperature biases the conditional probability to be uniform. When the temperature is about 0.6, the line begins to go up, and finally finds the optimum.

Temperature cooling rate is also an important factor that impacts the performance in ADA. Different cooling rates are compared in Figure 1(b), where the initial values are set to the means of the corresponding categories. Although all rates can provide comparatively good accuracies, a slower cooling rate will give a better result. When the cooling rate is too quick, such as $T_i = 0.1T_{i-1}$, only three steps of optimization are performed before $T = 0.001$. Therefore, the search for the optimal solution is insufficient before the conditional probabilities become hard, and such an optimization is easily trapped by local optima. In the annealing process, we also measure the time complexity when minimizing the cost at each temperature, as shown in Figure 1(c). We can see the algorithm is more time consuming at a high temperature, and speeds up as temperature decreases. At a given temperature, the SCG optimizer spends similar time when the temperature is low for different cooling rates. Hence, there is a tradeoff between the accuracy and time complexity. We should choose a faster cooling rate if we prefer an efficient algorithm; otherwise, we should select a slower cooling rate to get better solutions. No overfitting occurs in dADA in the WebKB collection.

In the inner product-based classifier, the regularization parameter ε controls the range of the parameter values, and the temperature T governs the level of randomness. The results for different regularization parameters are shown in Figure 1(d), where the starting point of each line is the accuracy of the initial parameter value. The initial parameters $\{\lambda_j\}$ are set to zero, which gives uniform conditional probabilities, an effect equivalent to that of infinite temperature. All the four lines have a similar shape. While gradually lowering the temperature, the accuracy increases until we get the best accuracy at the peak. Then the classifier begins to overfit, and accuracy drops. There is a relationship between the temperature T and the regularization parameter ε. When ε is large, the optimal solution is obtained at a higher temperature, and vice verse. The regularization parameter seems to only determine the temperature at which the maximum accuracy is reached. Therefore, no matter what the value of ε is, we can always find a good solution through annealing, which outperforms the logistic regression algorithm, whose accuracy is shown as the dashed line in the figure. Our model is equivalent to PLR when setting $T = 1$. Therefore, PLR is a special case of pADA. It is clear that the best regularization parameter of PLR is obtained by positioning the peak of the line at $T = 1.$, which is always a laborious task.

4.2 Newsgroups: Discussion Articles Collection

The Newsgroups data set is a comparatively large collection containing about 20000 articles evenly divided among 20 UseNet discussion groups. Many of the categories fall into confusable clusters; for example, five of them are comp.* discussion groups and three of them are religion. When tokenizing this collection, we skip the UseNet headers and subject line, and select 1000 features.



(a).The dADA accuracy with three initial values (1) means (2) random (3) zero are compared. The cooling rate is $T_i = 0.9T_{i-1}$. The final accuracy with different initial values converges when T is nearly zero.

(b).The dADA accuracy with four different cooling rates are compared. The initial values are set to the means of the categories. The slower cooling rate will get better accuracy.

(c).The dADA time complexity with different cooling rates, where the setting is the same as in (b). The cost decreases as temperature is low.

(d).The pADA accuracy based on different regularization parameters. The dashed line is the accuracy of the logistic regression without regularization.

Fig. 2. The experiments on the Newsgroup dataset. The x-axis is a logarithm scale.

For the distance-based classifier, its results are shown in Figure 2(a)-(c). They are similar to the results for WebKB. In Figure 2(a), since there are a total of 20 categories, the accuracy is about 5% for the random and zero initial values. LDA's accuracy is 55%. The distance-based classifier of ADA finally gets a much better result, i.e., 71% regardless of the initial values of the parameters. The conclusion that the optimal solution results from a slower cooling rate is more obvious in this experiment (Figure 2(b)). The accuracy from $T_i = 0.7T_{i-1}$ is nearly 10% higher than from $T_i = 0.1T_{i-1}$. Newsgroups is a dataset much larger than WebKB. Hence optimizer spends more time minimizing the cost function at a given temperature, as shown in Figure 2(c). As temperate lowers, it will spend less time. The time spent at a low temperature is only about one eighth of the time of a high temperature. For the inner product-based classifier, the result is shown in Figure 2(d), which is also very similar to the result for WebKB. The optimal classifier can obtain accuracy 75.5%, while that of logistic regression is 68.8%.

5 Conclusions

In this paper, we propose a novel classification method called annealed discriminant analysis (ADA). A probabilistic framework was constructed by randomization of the conditional probability, which is based on the principle of maximum entropy. The annealing process was introduced by controlling the Lagrange multiplier T based on the deterministic annealing approach, which is interpreted as gradually trading entropy of the associations for reduction of the cost function. While gradually lowering the temperature, the global optimum can be obtained independent of the choice of initial configuration. The distance-based classifier, an annealed version of linear discriminant analysis, outperforms the standard linear discriminant analysis. The inner-product based classifier, which can be seen as a generalized penalized logistic regression, provides the optimal solution, which is insensitive to the regularization parameter. The experiments demonstrate ADA's ability to provide substantial gains over existing methods.

References

1. Hofmann, T., Buhmann, J.: Pairwise data clustering by deterministic annealing. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 1–14
2. Miller, D., Rao, A.V., Rose, K., Gersho, A.: A global optimization technique for statistical classifier design. IEEE Transaction on Signal Processing **44** (1996) 3108–3122
3. Rao, A., Miller, D., Rose, K., Gersho, A.: A deterministic annealing approach for parsimonious design of piecewise regression models. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 159–173
4. Rose, K.: Deterministic annealing for clustering, compression, classification, regression, and related optimization problem. Proceedings of the IEEE **86** (1998) 2210–2239
5. Yuille, A.L., Stolortz, P., Utans, J.: Statistical physics, mixtures of distributions, and the *em* algorithm. Neural Computation **6** (1994) 334–340
6. Rose, K., Gurewitz, E., Fox, G.C.: Statistical mechanics and phase transitions in clustering. Physics Review Letter **65** (1990) 945–948
7. Zhang, T.: Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research **5** (2004) 1225–1251
8. Wahba, G.: Soft and hard classification by reproducing kernel Hilbert space methods. Proceedings of the National Academy of Sciences **99** (2002) 16524–16530
9. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions Pattern Analysis and Machine Intelligence **6** (1984) 721–741
10. Nabney, I.: Netlab: algorithms for pattern recognition. Springer-Verlag (2001)
11. Hastie, T., Tishiran, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag (2001)
12. McLachlan, G.J.: Discriminant analysis and statistical pattern recognition. John Wiley & Sons (1992)
13. Zhu, J., Hastie, T.: Classification of gene microarrays by penalized logistic regression. Biostatistics **5** (2004) 427–443