Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02), Vol. 2
Lipo Wang, Jagath C. Rajapakse, Kunihiko Fukushima,
Soo-Young Lee, and Xin Yao (Editors)

ADAPTIVE DETERMINISTIC ANNEALING FOR TWO APPLICATIONS: COMPETING SVR OF SWITCHING DYNAMICS AND TRAVELLING SALESMAN PROBLEMS.

Ming-Wei Chang and Chih-Jen Lin

National Taiwan University
Department of Computer Science
Taipei 106, Taiwan

Ruby C. Weng

National Chenechi University
Department of Statistics
Taipei 116, Taiwan

ABSTRACT

A deterministic annealing approach has been proposed to clustering by Rose et al. [10, 11] based on the maximum entropy principle. They obtain the association probabilities at a given average variance. The corresponding Lagrange multiplier is inversely related to the "temperature" and is used to control the annealing process. In this article we propose an adaptive annealing schedule where the Lagrange multiplier is treated as an unknown parameter and is estimated by an expectation-maximization step. This technique is applied to using support vector regression (SVR) for switching dynamics. We also give some preliminary results on Traveling Salesman Problems (TSP).

1. INTRODUCTION

Many application problems can be formulated as optimization problems by defining a cost function to be minimized. Most of these cost functions are nonconvex and have many local minima. Traditional gradient-based algorithms are ineffective because they tend to get trapped in local minima. An important stochastic method for avoiding local minima is simulated annealing, see Kirkpatrick, Gelatt, and Vecchi [7] and Laarhoven and Aarts [8] for an account of this method. However, the computational complexity of an implementation is very high. For example, in the area of image restoration, Geman and Geman [4] showed that, in theory, the global minimum can be achieved if the annealing schedule obeys $T \propto 1/\log n$, where T is the temperature and n is the number of the current iteration; Hajek [5] and Szu and Hartley [12] also derive sufficient conditions for simulated annealing algorithms to statistically converge to a global minimum, and these conditions require the annealing schedules being "slow enough."

Unlike simulated annealing where random moves are made on the given energy surface, deterministic annealing (DA) can be viewed as incorporating the "randomness" into the energy function by extracting properties of the macroscopic system from microscopic averages. In this approach, an effective energy function is obtained and is deterministically optimized at each temperature sequentially, starting from high temperature and going down. This approach was adopted by Durbin and Willshaw [3] and Durbin, Szeliski, and Yuille [2] for the traveling salesman problem, by Rose, Gurewitz, and Fox [10, 11] for clustering problems, by Hofmann, Puzicha, and Buhmann [6] for unsupervised texture segmentation, and among others.

The idea of DA is applying the maximum entropy principle to obtain the association probabilities at a given average variance. The corresponding Lagrange multiplier is inversely related to the "temperature" and is used to control the annealing process. In this process, as the temperature is gradually decreased, the system undergoes a sequence of "phase transitions."

Recently, Pawelzik, Kohlmorgen, and Müller [9] propose to deterministically anneal the competition of neural networks to solve a time series segmentation and identification problem. Under a similar framework, Chang, Lin, and Weng [1] exploit the use of support vector regression (SVR) instead of neural networks. Recently, support vector machines (SVM) [13] have been a promising method for data classification and regression. The good generalization error is obtained by using the maximum margin. Unlike earlier DA work, the authors of [1] treat β as an unknown parameter and adaptively update it by an expectation-maximization (EM) step. We shall call it *adaptive deterministic annealing* (ADA). This paper aims to give a thorough discussion of ADA. In Section 2, we review DA and introduce ADA. The properties of ADA are addressed. In Section 3, we illustrate the performance of our approach by considering the time series segmentation problem in [9]. We explain in detail how competing SVRs fit into the ADA framework. We then in Section 4 apply ADA to the elastic net (EN) approach for the traveling salesman problem (TSP) by Durbin and Willshaw [3]. Both examples exhibit how ADA considerably speeds up the annealing process while maintains high accuracy.

2. ADAPTIVE ANNEALING AND ITS PROPERTIES

2.1. Review of DA

Suppose we have l data points, $x_1, ..., x_l$, coming from m clusters, where x_t is a q-dimensional vector $[x_{t1}, ..., x_{tq}]^T$ for $t = 1, ..., l$. Let $P(x \in C_j)$ be the probability that x belongs to the cluster C_j and $d(x, y)$ the cost for representing data point x by the parameter vector y. Assume the set of representatives $Y = \{y_j\}$ is given, where y_j specifies the cluster C_j. Then the expected energy is

$$E = \sum_t \sum_j P(x_t \in C_j) d(x_{t,}, y_j). \tag{1}$$

Under the expectation constraint, the probability distributions that maximize the entropy

$$H = -\sum_t \sum_j P(x_t \in C_j) \log P(x_t \in C_j)$$

are Gibbs distributions,

$$P(x \in C_j) = \frac{\exp(-\beta d(x, y_j))}{\sum_k \exp(-\beta d(x, y_k))}. \tag{2}$$

If we use l_2-norm as the cost function, $d(x,y) = |x-y|^2$, the association probability (2) corresponds to a Gaussian distribution. If we take l_1-norm, then it becomes Laplace distribution (also called double-exponential).

The parameter β is the Lagrange multiplier determined by the given value of E in (1). The Lagrangian function is $H - \beta E$. Interestingly, we can view deterministic annealing as an static physical system. The formulation is:

$$F = E - TH, \tag{3}$$

where F is the free energy and T is the constant temperature. According to the principle of minimal free energy, F should be minimized to achieve the equilibrium.

We can view (3) as a Lagrangian function and T is the multiplier. In other words, we minimize the energy under some "randomness," and then we lower the temperature to put more weight on the original energy function. When temperature downs to zero, (3) will become (1) and then we get the global minimum.

After substituting T by $\frac{1}{\beta}$ and $P(x \in C_j)$ by the value in (2), the free energy (3) is

$$F(Y) = -\frac{1}{\beta}\sum_t \log[\sum_k e^{-\beta d(x_t, y_k)}]. \tag{4}$$

The set of vectors Y which optimizes the free energy satisfies

$$\frac{\partial}{\partial y_j} F = 0, \ \forall j, \tag{5}$$

where this is a shorthand notation for differentiation with respect to each component separately. Differentiating (4) we obtain

$$\sum_t P(x_t \in C_j) \frac{\partial}{\partial y_j} d(x_t, y_j) = 0. \tag{6}$$

In the case of l_2-norm, (5) is simplified as

$$y_j = \frac{\sum_t x_t P(x_t \in C_j)}{\sum_t P(x_t \in C_j)}. \tag{7}$$

In practice we can only approximately minimize F. Note that when $P(x_t \in C_j)$ is fixed, minimizing F is the same as minimizing E. At $\beta = 0$, we have a single minimium for F (which is also the global minimum). At positive β we may have several local minima. The concept of annealing can be viewed as tracking the minimum, starting at the global minimum and gradually increasing β. This avoids arbitrary local minima depending on the initialization of the iterations. The procedure is as follows: (i) Initialize β to be a small number near zero. (ii) Solve y_j from (??). (iii) Calculate the expected cluster membership by (2). (iv) Increase β and repeat (ii)-(iii).

At $\beta = 0, y_j = \bar{x} = \sum_t x_t/m$, so we have a single cluster. Then at some positive β the cluster will split into smaller cluster, and will thus undergo a phase transition. Rose et al. [10, 11] compute the approximate critical β for the first transition. Let C_{xx} be the covariance matrix $\sum(x_t - \bar{x})(x_t - \bar{x})'/l$, and $\lambda_{\max}$ the maximum eigenvalue of C_{xx}. By a careful examination of the differentiation of (1), they show that the first transition occurs at $\beta_c \approx 1/(2\lambda_{\max})$. As long as the intercluster influences can be neglected, this derivation will hold for the following phase transitions. Knowing to predict the next critical β may allow us to accelerate the process between phase transitions, while being more careful during the transitions.

2.2. Introducing ADA

Unlike DA that increase β steadily, we propose to increase it based on the results of the previous iteration. Through out this section we denote the predicted errors as $e_j^t = x_t - \hat{y}_j$. ¿From (2) β can be viewed as an unknown parameter that characterizes some property about e. Thus, at each iteration we suggest to update β by maximizing the conditional likelihood function given current parameter estimates, which we believe can best illustrates e. More specifically, denote the representatives at the g-th iteration as $\hat{y}^{(g)} = (\hat{y}_1^{(g)}, ..., \hat{y}_m^{(g)})'$. Given $\hat{y}^{(g)}$, we suppose that x_t follows a mixture of distributions p_i with mean $\hat{y}_i^{(g)}$, unknown parameter τ, and mixing coefficients a_i. Then, the density for x_t is

$$p(x_t|y^{(g)}, \tau) \quad = \quad \sum_{i=1}^{m} a_i p_i(x_t|y^{(g)}, \tau).$$

If we introduce an additional variable S_t such that $S_t = i$ when $x_t \in C_i$, then the density function can be simplied as

$$p(x_t, s_t|y^{(g)}, \tau) = a_{s_t} p_{s_t}(x_t|y^{(g)}, \tau). \tag{8}$$

As S_t are latent (unobservable) variables, we use an expectation-maximization (EM) step to compute the maximum likelihood estimate. The E-step calculates the expectation of the conditional log-likilihood function given $(X, y^{(g)}, \tau^{(g)})$, which is

$$\begin{aligned}
& Q(\tau; \tau^{(g)}) \\
= \quad & E[\log p(X, S|y^{(g)}, \tau)||X, y^{(g)}, \tau^{(g)}] \\
= \quad & \sum_{t=1}^{l} E\{\log[\alpha_{s_t} p_{s_t}(x_t|y^{(g)}, \tau)]||X_t, y^{(g)}, \tau^{(g)}\} \\
= \quad & \sum_{t=1}^{l}\sum_{i=1}^{m}\{(\log\alpha_i) p_i(x_t|y^{(g)}, \tau) p_i^{t(g)} \\
= \quad & \sum_{t=1}^{l}\sum_{i=1}^{m}\{(\log\alpha_i) p_i^{t(g)}\} + \sum_{t=1}^{l}\sum_{i=1}^{m}\{[\log p_i(x_t|y_t, \tau_i)] p_i^{t(g)}\},
\end{aligned}$$

where $p_i^{t(g)} = P(S_t = i|X_t, y^{(g)}, \tau^{(g)})$, and the third equality follows because when conditioning on $(X, y^{(g)}, \tau^{(g)})$

$$\alpha_{s_t} p_{s_t}(x_t|y^{(g)}, \tau) = \alpha_i p_i(x_t|y^{(g)}, \tau)$$

with probability $p_i^{t(g)}$, for $i = 1, ..., m$. The M-step finds maximizer of the expectation,

$$\tau^{(g+1)} = \arg\max_{\tau} Q(\tau; \tau^{(g)}). \tag{9}$$

We suppose that the distribution of the predicted errors at e_j^t to be symmetric about zero. Suppose also that p_i belongs to exponential power family of density functions with location parameter μ and scale parameter λ; that is, $p_i(z|\mu, \lambda) \propto \exp(-|z-\mu|^r/\lambda)$. It can be shown that

$$p_i(z|\mu, \lambda) = k\tau^{-1/r} \exp(-|z-\mu|^r/\lambda),$$

where k is a normalizing constant and it does not depend on λ. Particularly, for $r = 1$, it is a Laplace distribution; for $r = 2$, it is Gaussian. If z is a q-dimensional vector, we have

$$p_i(z|\mu, \lambda) = k^q \tau^{-q/r} \exp(-||z-\mu||^r/\lambda),$$

where

$$\|z-\mu\|^r = \sum_{j=1}^{q} |z_j - \mu_j|^r$$

So the solution to (8) is

$$\tau^{(g+1)} = \frac{\sum_{t=1}^{l} \sum_{i=1}^{m} \|e_i^{t(g)}\|^r p_i^{t(g)}}{lq/r}. \qquad (10)$$

Now we write the posterior probability $p_i^{t(g)}$ as

$$\begin{aligned} p_i^{t(g)} &= \frac{P(S_t = i, X_t|y^{(g)}, \tau^{(g)})}{\sum_j P(S_t = j, X_t|y^{(g)}, \tau^{(g)})} \\ &= \frac{a_i p_i(x_t|y^{(g)}, \tau^{(g)})}{\sum_j a_j p_j(x_t|y^{(g)}, \tau^{(g)})} \\ &= \frac{a_i \exp(-\|e_i^{t(g)}\|^r/\tau)}{\sum_j a_j \exp(-\|e_j^{t(g)}\|^r/\tau)}. \qquad (11) \end{aligned}$$

By comparing (2) and (10), if we take the cost function as $d(x, y) = \|x - y\|^r$, then β in (2) plays about the same role as $1/\tau$ in (10). So after obtaining $\tau^{(g+1)}$, we suggest to update β by

$$\beta^{(g+1)} = 1/\tau^{(g+1)}. \qquad (12)$$

The procedure is suggested as follows: (i) The initial value p_i^t is randomly assigned to be 1 or 0, subject to $\sum_{i=1}^{m} p_i^t = 1$. (ii) Calculate y_j by (6). (iii) Update β by (9) and (11) and re-calculate p_i^t by (10). (iv) Repeat (ii) and (iii) until reaching some stopping criteria.

We use $r = 2$ to approximate the first β updated by our method. Note that at the first iteration, $\hat{y}_i = [\hat{y}_{i1}, ..., \hat{y}_{iq}]^T \approx \bar{x} = [\bar{x}_{.1}, ..., \bar{x}_{.q}]^T$ for $i = 1, ..., m$, where $\bar{x}_j = \sum_{t=1}^{l} x_{tj}/l$. Together with (9), we obtain

$$\begin{aligned} \tau &= \frac{2}{lq} \sum_{t=1}^{l} \sum_{i=1}^{m} [\sum_{j=1}^{q} (x_{tj} - \hat{y}_{ij})^2] p_i^t \\ &\approx \frac{2}{lq} \sum_{t=1}^{l} \sum_{j=1}^{q} (x_{tj} - \bar{x}_j)^2 \\ &= \frac{2}{q} \cdot \text{trace}(C_{xx}) \\ &= \frac{2}{q} \cdot \text{sum of eigenvalues of } C_{xx} \\ &\geq \frac{2}{q} \lambda_{\max}. \end{aligned}$$

If $q = 1$, then C_{xx} is 1×1 and $\beta = 1/\tau \approx 1/(2\lambda_{\max})$. In practice, as the distribution of e_j^t is unknown, we usually try $r = 1$ or 2 and multiply $\beta^{(g+1)}$ by a factor of 2 to 5.

3. SUPPORT VECTOR REGRESSION FOR TIME SERIES SEGMENTATION

Pawelzik, Kohlmorgen, and Müller [9] use an architecture consisting of competing neural networks to segment data streams originating from different unknown sources which alternate in time. For unique segmentation, each data point must be assigned to only one predictor. This can be achieved by considering only the respective best performing predictor. However, using hard competition right from the beginning usually gets stuck in local minima. So they adopt deterministic annealing to gradually increase the degree of competition. Chang, Lin, and Weng [1] propose to test the same problem using competing support vector machines and the adaptive annealing method. Given $(x_t, y_t), t = 1, ..., l$, we have

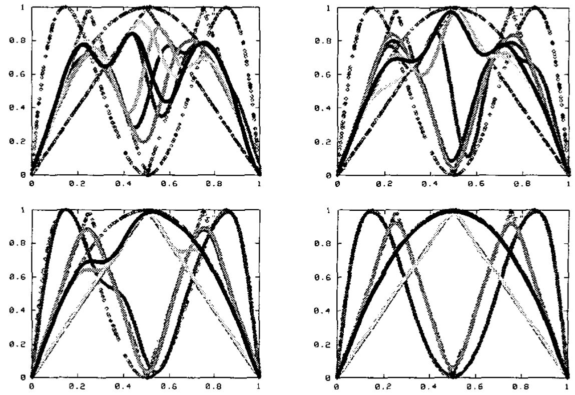

Figure 1: First four iterations (data without noise)

$$y_t = f_{r_t}(x_t), \qquad (13)$$

where $r_t \in 1, ..., m$. The task is to find out r_t which indicates that (x_t, y_t) is with which function (series).

Now (x_t, y_t) is like x_t in (1) and if we take l_2-norm, $d(x_t, y_j)$ becomes

$$(y_t - f_j(x_t))^2. \qquad (14)$$

f_j is the approximate function, which now is the representative of the jth cluster. Then, the Gibbs distributions in (2) becomes

$$P((x_t, y_t) \in C_j) = \frac{\exp(-\beta(y_t - f_j(x_t))^2)}{\sum_j \exp(-\beta(y_t - f_j(x_t))^2)}. \qquad (15)$$

In this section, we denote $P((x_t, y_t) \in C_i)$ as p_i^t . Unfortunately, such a setting is not enough for segmenting series. In [1], following from [9], the authors assume low switching rate. That is, data before and after any given time point t are likely to come from the same series. Therefore, instead of using (14), after new $\hat{f}_i$ are obtained, they update p_i^t by

$$p_i^t = \frac{\exp(-\beta \sum_{\delta=-\Delta}^{\Delta} (e_i^{t-\delta})^2)}{\sum_{j=1}^{m} \exp(-\beta \sum_{\delta=-\Delta}^{\Delta} (e_j^{t-\delta})^2)}, \qquad (16)$$

where

$$e_i^t = y_t - \hat{f}_i(x_t). \qquad (17)$$

We consider $f_i(x)$ has the following form

$$f_i(x) = w_i^T \phi(x) + b, \qquad (18)$$

where x is mapped to a higher dimensional space by ϕ. We slightly change the energy function to be

$$E = \sum_t \sum_i p_i^t d((x_t, y_t), f_i) + \frac{1}{2C} \sum_i w_i^T w_i. \qquad (19)$$

If we use

$$d((x_t, y_t), f_i) = |y_t - (w_i^T \phi(x_t) + b_i)|_\epsilon, \qquad (20)$$

where $|\cdot|_\epsilon$ is the ϵ-insensitive loss function, then minizeing the free energy with p_i^t fixed is

$$\min_{w_i, b_i} \quad \sum_i \left(\frac{1}{2} w_i^T w_i + C \sum_{t=1}^{l} p_i^t d((x_t, y_t), f_i)\right). \qquad (21)$$

(20) can be separated to m independent problems: For $i = 1, \dots, m$, we solve

$$\begin{aligned} \min_{w_i, b_i, \xi, \xi^*} \quad & \frac{1}{2} w_i^T w_i + C \sum_{t=1}^{l} p_i^t (\xi_i^t + \xi_i^{t,*}) \\ \text{s.t.} \quad & -\epsilon - \xi_i^{t,*} \leq y_t - (w_i^T \phi(x_t) + b) \leq \epsilon + \xi_i^t, \\ & \xi_i^t \geq 0, \xi_i^{t,*} \geq 0, t = 1, \dots, l, \end{aligned}$$

which is a modification of the standard support vector regression. $\frac{1}{2} \sum_i w_i^T w_i$ can be treated as a "regularization term" which avoids overfitting. See [1] for further discussion.

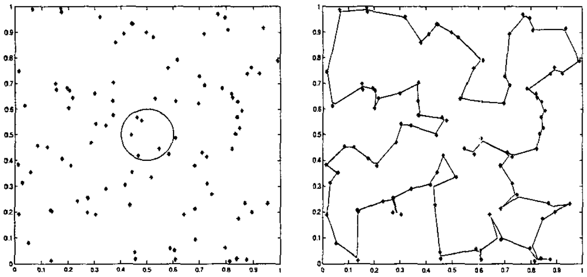

Figure 2: Results using deterministic annealing.

A minor problem of the procedure in [1] is that (21) and (15) cannot fit into the deterministic annealing framework. In the following we propose a further modification which solves this problem. By considering,

$$d((x_t, y_t), f_i) = \sum_{\delta=-\Delta}^{\Delta} |e_i^{t-\delta}|, \qquad (22)$$

we define the excepted energy function as follows

$$E = \sum_t \sum_i p_i^t \sum_{\delta=-\Delta}^{\Delta} |e_i^{t-\delta}| + \frac{1}{2C} \sum_i w_i^T w_i. \qquad (23)$$

Following (2),

$$p_i^t = \frac{\exp\left(-\beta \sum_{\delta=-\Delta}^{\Delta} |e_i^{t-\delta}|\right)}{\sum_{j=1}^{m} \exp\left(-\beta \sum_{\delta=-\Delta}^{\Delta} |e_j^{t-\delta}|\right)}. \qquad (24)$$

Then we minimize E under fixed p_i^t,

$$\begin{aligned} \min_{w_i, b_i, \xi, \xi^*} \quad & \frac{1}{2} w_i^T w_i + \frac{C}{2\Delta + 1} \sum_{t=1}^{l} p_i^t \sum_{\delta=-\Delta}^{\Delta} (\xi_i^{t+\delta} + \xi_i^{t+\delta,*}) \\ \text{s.t.} \quad & -\epsilon - \xi_i^{t,*} \leq y_t - (w_i^T \phi(x_t) + b) \leq \epsilon + \xi_i^t, \\ & \xi_i^t \geq 0, \xi_i^{t,*} \geq 0, t = 1, \dots, l. \end{aligned}$$

Note that C is the penalty parameter of SVR given by users. Hence, we can denote it as $C/(2\Delta + 1)$. (25) is the same as

$$\begin{aligned} \min_{w_i, b_i, \xi, \xi^*} \quad & \frac{1}{2} w_i^T w_i + C \sum_{t=1}^{l} \left(\sum_{\delta=-\Delta}^{\Delta} \frac{p_i^{t+\delta}}{2\Delta + 1}\right)(\xi_i^t + \xi_i^{t,*}) \\ \text{s.t.} \quad & -\epsilon - \xi_i^{t,*} \leq y_t - (w_i^T \phi(x_t) + b) \leq \epsilon + \xi_i^t, \\ & \xi_i^t \geq 0, \xi_i^{t,*} \geq 0, t = 1, \dots, l, \end{aligned}$$

which is also a modification of the standard support vector regression. Note that we assume $p_i^t = 0$ if t is not in the range of $\{1, \dots, l\}$. Now we can clearly see that $\sum_{\delta=-\Delta}^{\Delta} p_i^{t+\delta}/(2\Delta + 1)$ is the average of a moving window. By doing so we feel that the choice of C is independent of Δ.

Next, we illustrate the proposed procedure by a simple example. We consider all $x_t \in [0, 1]$ and four different functions: $f_1(x) = 4x(1 - x)$, $f_2(x) = 2x$ if $x \in [0, 0.5)$ and $2(1 - x)$ if $x \in [0.5, 1]$, $f_3(x) = f_1(f_1(x))$, and $f_4 = f_2(f_2(x))$. It is easily seen that all these functions map x from $[0, 1]$ to $[0, 1]$.

Figure 1 shows the results of the first four iterations using our approach. In this example the approximation results of critical values by Rose et al. [10, 11] can not be applied to speed up the algorithm. But our approach does work well. The adaptive method also works for the Mackey-Glass example, but is not reported here.

4. TRAVELING SALESMAN PROBLEM

In the clustering problem described in Section 1, if we put in enough representatives and let $\beta \to \infty$, then each data point itself will become a cluster. As an example, we consider the traveling salesman problem. It is stated as follows: Given the positions of N cities, find the shortest closed path that passes through all of them. Durbin and Willshaw [3] proposed an elastic net (EN) method to solve it. The EN starts with a set of representing points that form a small circle near the centroid of the cities, then gradually elongates the points non-uniformly to pass eventually near all the cities and thus defined a tour around them. As in Durbin and Willshaw [3], suppose that the sum of squared distance between consecutive cities on the path is to be minimized. Let x_i denote the coordinates of city i and $y_j, j = 1, \dots, M$ the coordinates of point j on the path. Then the free energy function is defined as

$$F = -\alpha K \sum_i \ln \sum_j \phi(|x_i - y_j|, K) + \gamma \sum_j |y_{j+1} - y_j|^2,$$

where $\phi(d, K) = \exp(-d^2/2K^2)$. To minimize F, we consider the steepest descent direction:

$$\begin{aligned} \Delta y_j &= -K \partial F / \partial y_j \\ &= \alpha \sum_i p_j^i (x_i - y_j) + 2\gamma K (y_{j+1} - 2y_j + y_{j-1}), \end{aligned}$$

where

$$p_j^i = \phi(|x_i - y_j|, K) / \sum_s \phi(|x_i - y_s|, K).$$

It is difficult to find the minimum of F, so when p_i^t is fixed, we intend to decrease F by modifying y_j with several steps Δy_j . It is easily seen that the β in our previous sections is the same as $1/(2K^2)$ in this setting. In an example of $N = 100$ cities randomly distributed in the unit square, parameter values in [3] were: $\alpha = 0.2$, $\gamma = 2.0$, and the initial number of representatives is $M = 2.5N$, the initial value of K was 0.2, and was reduced by 1% every 25 updates of y_j. In the end, K was reduced to 0.01 in 7,000 updates. Here, we conduct a similar experiment. We take $N = 100$, $M = 1.4N$, $\alpha = 1.0$, and $\gamma = 2.0$. For deterministic annealing, the initial β is set to be $1/200$ and is increased by $1/200$ for every 10 updates to a final value of 2.5. Figure 2 shows the results using deterministic annealing. For the adaptive method, the β is adapted by (11). For each β, 10 steps of Δy_j are taken. The results of the 1, 3, 8, 10, 12, and 14 iterations are shown in Figure 3. As the iterations go on, we remove some representatives that have little contribution to the configuration. In the end the tour length is about 7.65, very close to the value reported in [3]. Comparing to thousands of steps on updating Δy_j using DA, the proposed approach, with less than 200 steps, is much faster.

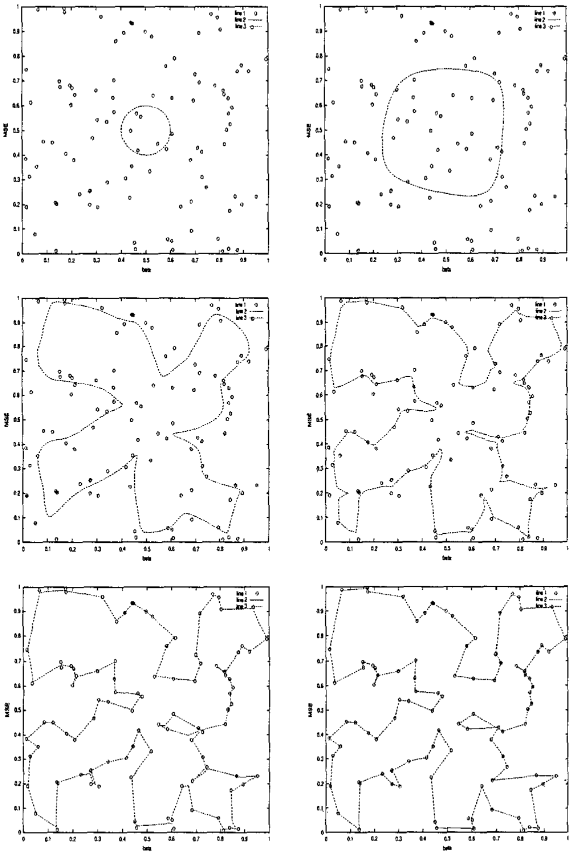

Figure 3: Results of the 1, 3, 8, 10, 12, 14 iterations using adaptive method.

5. REFERENCES

[1] M.-W. Chang, C.-J. Lin, and R. C. Weng. Analysis of switching dynamics with competing support vector machines. In *Proceedings of IJCNN*, 2002.

[2] R. Durbin, R. Szeliski, and A. L. Yuille. An analysis of the elastic net approach to the travelling salesman problem. *Neural Computation*, 1:348–358, 1989.

[3] R. Durbin and D. Willshaw. An analogue approach of the travelling salesman problem using an elastic net method. *Nature*, 326:689–691, 1987.

[4] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[5] B. Hajek. A tutorial survey of theory and applications of simulated annealing. In *24th IEEE Conf. Decision Contr.*, pages 755–760, 1985.

[6] T. Hofmann, J. Puzicha, and J. M. Buhmann. Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):803–818, 1998.

[7] S. Kirkpatrick, C. D. G. Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[8] P. Laarhoven and E. Aarts. *Simulated annealing: Theory and applications.* Reidel Pub., Dordrecht, Holland, 1987.

[9] K. Pawelzik, J. Kohlmorgen, and K.-R. Müller. Annealed competition of experts for a segmentation and classification of switching dynamics. *Neural Computation*, 8(2):340–356, 1996.

[10] K. Rose, F. Gurewitz, and G. Fox. Statistical mechanics and phase transitions in clustering. *Physical Rev. Letters*, 65(8):945–948, 1990.

[11] K. Rose, F. Gurewitz, and G. Fox. Vector quantization by deterministic annealing. *IEEE Trans. Inform. Theory*, 38:1249–1258, 1992.

[12] H. Szu and R. Hartley. Fast simulated annealing. *Physics Letters A*, 122:157–162, 1987.

[13] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.