RECURSIVE BAYESIAN REGRESSION MODELING AND LEARNING

Jen-Tzung Chien and Jung-Chun Chen

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC
{chien, jcchen}@chien.csie.ncku.edu.tw

ABSTRACT

This paper presents a new Bayesian regression and learning algorithm for adaptive pattern classification. Our aim is to continuously update regression parameters to meet nonstationary environments for real-world applications. Here, a kernel regression model is used to represent two-class data. The initial regression parameters are estimated by maximizing the likelihood of training data. To activate online learning, we properly express the randomness of regression parameters as a *conjugate prior*, which is a normal-gamma distribution. When new adaptation data are enrolled, we can accumulate *sufficient statistics* and come up with a new normal-gamma distribution as the posterior distribution. We therefore exploit a recursive Bayesian algorithm for online regression and learning. Regression parameters are incrementally adapted to the newest environments. Robustness of classification rule is assured using online regression parameters. In the experiments on face recognition, the proposed regression algorithm outperforms support vector machine and relevance vector machine for different numbers of adaptation data.

Index Terms—Recursive Bayesian learning, kernel regression, incremental adaptation, support vector machine, pattern recognition

1. INTRODUCTION

Due to the power of model generalization, support vector machine (SVM) [13] has been attracting many researchers working on the related issues. Continuing improvements in SVM have led into many successful applications in pattern recognition, e.g. information retrieval [10], speech recognition [11], and face recognition [7]. However, in real-world applications, such generalization is not sufficient to meet continuously changing environments. For example, face recognition performance is always degraded when the environmental conditions of illumination, facial expression, and pose angle are mismatch with those in training data. We should build adaptation/learning mechanism to overcome mismatch problem. Especially, in nonstationary environments, testing conditions are varied all the time. How to build online adaptation for SVM or other models becomes important for robust pattern recognition [4]. In this study, we are developing online adaptation algorithm from Bayesian learning viewpoint. In the literature, Bayesian interpretation of SVM has been discussed in [5][6]. Bayesian support vector regression with an evidence framework was proposed to connect the relations to MacKay's Bayesian framework [8]. Nevertheless, it was necessary to compute the integral in evidence framework. Some approximations were engaged so that no exact Bayesian explanation for SVM was available. Although SVM could be interpreted by Bayesian theory, these methods were not developed for Bayesian learning in presence of nonstationary environments.

For these results, we present a new recursive Bayesian (RB) regression framework for general pattern classification. This framework can fulfill the spirit of SVM, namely maximization of margin and minimization of classification errors. Attractively, we adopt the conjugate prior distribution to express the perturbation of regression parameters. When adaptation data are collected, we can produce a posterior distribution, which belongs to the same distribution family as prior function [2]. This reproducible prior/posterior pair activates an online learning strategy for kernel regression model. This RB method can be also simplified as a batch learning that only one learning epoch is executed. We investigate the performance of online regression model on FERET facial database containing various conditions of lighting, expression and orientation. In what follows, we begin with a survey of regression models for pattern classification. Subsequently, we construct the recursive Bayesian approach to kernel regression model and explain its relations to previous methods. Finally, we demonstrate the performance of proposed method and draw some conclusions.

2. RELATED WORKS

We are introducing Bayesian kernel regression methods. Some notations should be explained. Given input data $\mathbf{x} \in \Re^m$, we transform it to high dimensional space $\mathbf{z} \in \Re^d$, $d > m$, using function $\varphi(\mathbf{x})$. We collect a set of input-output training samples $D = \{(\mathbf{z}_i, y_i)\}_{i=1}^n$ for supervised training. Kernel function is calculated by dot product of two samples $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j) = \mathbf{z}_i \cdot \mathbf{z}_j$. Considering a two-class pattern classification problem, output y_i represents class label, e.g. $y_i = +1$ for class 1 and $y_i = -1$ for class 2.

2.1. Bayesian Support Vector Regression

Using SVM, we are seeking the optimal hyperplane $f(\mathbf{w},\mathbf{z}) = \mathbf{w}^T\mathbf{z} + b$, which has the maximum margin between support vectors of two classes. Vapnik [13] also extended SVM to kernel-based support vector regression (SVR). In SVR, training samples of a class are represented by

$$y_i = \mathbf{w}^T\mathbf{z}_i + b + \varepsilon + \xi_i . \quad (1)$$

This ε insensitive loss function [13] was presented to obtain the sparseness property of SVM. Hence, we have $\xi_i, \xi_i^* = \left|y_i - f(\mathbf{x}_i,\mathbf{w})\right|_\varepsilon$ as a modeling error with value of ε being margin width. ε is empirically defined. Using this loss function, the estimation accuracy of outliers will not hurt too much. The optimal parameters of $\mathbf{w}$ and b are obtained according to the least square error criterion

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) . \quad (2)$$

How to automatically determine regularization parameter C for various datasets becomes a crucial issue. Kwok [5] illustrated the relationship between SVM and MacKay's Bayesian framework. The hyperparameters of SVM were estimated by maximizing the evidence.

In SVR model [6], the parameter $\mathbf{w}$ was assumed to be random and $C = \beta/\alpha$. Meaningfully, the hyperparameter α controls the model complexity. The hyperparameter β controls the training error. The prior distribution of $\mathbf{w}$ is Gaussian distributed

$$p(\mathbf{w}|\alpha) \propto \exp(-\frac{\alpha}{2}\cdot\|\mathbf{w}\|^2) . \quad (3)$$

Assuming training samples are i.i.d., the output distribution of training samples D is [5]

$$p(D|\mathbf{w},\beta) = \prod_{i=1}^{n} p(\mathbf{z}_i, y_i|\mathbf{w},\beta) = \prod_{i=1}^{n} p(y_i|\mathbf{z}_i,\mathbf{w},\beta)p(\mathbf{z}_i) , \quad (4)$$

where an exponential distribution is specified as

$$p(y_i|\mathbf{z}_i,\mathbf{w},\beta) = \frac{\beta\exp(-\beta|\xi_i|_\varepsilon)}{2(1+\varepsilon\beta)} . \quad (5)$$

The posterior probability of $\mathbf{w}$ can be formulated by

$$p(\mathbf{w}|D,\alpha,\beta) = \frac{p(D|\mathbf{w},\beta)p(\mathbf{w}|\alpha)}{\int p(D|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}} . \quad (6)$$

The optimal parameter of $\mathbf{w}$ is calculated by maximizing this posterior probability. Optimal MAP estimate $\mathbf{w}_{\mathrm{MAP}}$ is used to approximate the evidence in (6). Optimal values of α and β are then obtained by iterating the process of finding $\mathbf{w}_{\mathrm{MAP}}$ and maximizing the evidence. Parameter b is determined correspondingly. The integral in evidence is done in accordance with MacKay's evidence framework [8]. In [1], SVR was regarded as a classification problem in the dual space. No model learning was concerned in these works.

2.2. Relevance Vector Machine

Accordingly, Tipping [12] presented Bayesian learning via a relevance vector machine (RVM). Different from Bayesian SVR [6], RVM assumed that parameters $\boldsymbol{\alpha}$ and β are random. The training samples are represented as

$$y_j = \sum_{i=1}^{n} w_i K(\mathbf{x}_j,\mathbf{x}_i) + b + \xi_j . \quad (7)$$

where the relevance is revealed by kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$. Let $\tilde{\mathbf{w}} = [\mathbf{w}^T, b]^T$ denote model parameters. A normal distribution with adjustable variance $\boldsymbol{\alpha}$ is assumed for $\tilde{\mathbf{w}}$. The gamma distribution over $\boldsymbol{\alpha}$ with hyperparameter λ_α is given by

$$p(\boldsymbol{\alpha}) = \prod_{i=0}^{n}\mathrm{Gamma}(\alpha_i|\lambda_\alpha) . \quad (8)$$

A zero-mean Gaussian prior distribution over $\tilde{\mathbf{w}}$ is given as

$$p(\tilde{\mathbf{w}}|\boldsymbol{\alpha}) = N(b|0,\alpha_0^{-1})\cdot\prod_{i=1}^{n}N(w_i|0,\alpha_i^{-1}) . \quad (9)$$

The likelihood function of all the data conditional on unknown parameters is expressed by

$$p(\mathbf{y}|\tilde{\mathbf{w}},\beta) = \frac{(\beta)^{n/2}}{(2\pi)^{n/2}}\left\{\exp\left[-\frac{\beta}{2}(\mathbf{y}-\boldsymbol{\Phi}\tilde{\mathbf{w}})^T(\mathbf{y}-\boldsymbol{\Phi}\tilde{\mathbf{w}})\right]\right\} \quad (10)$$

where $\boldsymbol{\Phi} = [\phi(\mathbf{x}_1),\ldots,\phi(\mathbf{x}_n)]^T$ is the $n\times(n+1)$ matrix and $\phi(\mathbf{x}_i) = [K(\mathbf{x}_i,\mathbf{x}_1),\ldots,K(\mathbf{x}_i,\mathbf{x}_n),1]^T$. Parameter β is a precision in density (10). The gamma distribution over β is also engaged in

$$p(\beta) = \mathrm{Gamma}(\beta|\lambda_\beta) . \quad (11)$$

These priors were assumed to be non-informative. Using (6), (9) and (10), we obtain the posterior distribution over $\tilde{\mathbf{w}}$ as a Gaussian distribution

$$p(\tilde{\mathbf{w}}|\mathbf{y},\boldsymbol{\alpha},\beta) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{(n+1)/2}}\left\{\exp\left[-\frac{1}{2}(\tilde{\mathbf{w}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{w}}-\boldsymbol{\mu})\right]\right\} , \quad (12)$$

where the mean vector is $\boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{y}$, the covariance matrix is $\boldsymbol{\Sigma} = (\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{\alpha}\mathbf{I})^{-1}$, and $\mathbf{I}$ is the $(n+1)\times(n+1)$ identity matrix. We have the result of $\mathbf{w}_{\mathrm{MAP}} = \boldsymbol{\mu}$. Optimal $\boldsymbol{\alpha}$ and β are obtained by MacKay's evidence framework [8]. Finally, we can classify a test pattern according to likelihood function of (10).

3. RECURSIVE BAYESIAN REGRESSION

Basically, Bayesian SVR [5][6] provided Bayesian solution to SVM and extended SVM paradigm to deal with regression problem. Instead of merging support vectors, RVM [12] took relevance vectors into account and derive the predictive posterior distribution to activate Bayesian learning. Prior densities of parameters $\tilde{\mathbf{w}}$ and their hyperparameters $\boldsymbol{\alpha}, \beta$ were defined. It was necessary to approximate the multivariate integral of evidence in MacKay's framework [8]. In this study, we focus on developing Bayesian and incremental learning strategy for kernel regression model. Dummy variable is introduced in regression model so as to fit in with the maximum margin criterion adopted in SVM or SVR.

3.1. Regression Model Estimation

To follow up the objective of maximum margin in SVM, the dummy variable $c_i \in \{0,1\}$ is merged in regression model for each data pair [9]. This dummy label is used to describe the supervision of input pattern. Similar to SVR in (1), the training samples are now represented as

$$y_i = \mathbf{w}^T \mathbf{z}_i + b + c_i \varepsilon + \xi_i , \tag{13}$$

where ε can be regarded as the margin between two classes. The model parameters turn out to be $\tilde{\mathbf{w}} = [\mathbf{w}, b, \varepsilon]^T$ and the input pattern can be extended to $\tilde{\mathbf{z}}_i = [\mathbf{z}_i^T, 1, c_i]^T$. We can find the least-square solution to regression parameters $\tilde{\mathbf{w}}$ by minimizing

$$\sum_{i=1}^{n} \xi_i^2 = \boldsymbol{\xi}^T \boldsymbol{\xi} = (\mathbf{y} - \tilde{\mathbf{Z}}\tilde{\mathbf{w}})^T (\mathbf{y} - \tilde{\mathbf{Z}}\tilde{\mathbf{w}}) , \tag{14}$$

where $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n]^T$. Reducing this expected error is able to achieve maximum margin and minimum classification error. The least-squares estimator is built by $\hat{\mathbf{w}} = \tilde{\mathbf{Z}}^T \tilde{\boldsymbol{\Phi}}^{-1} \mathbf{y}$ where $\tilde{\boldsymbol{\Phi}} = [\tilde{\boldsymbol{\Phi}}_1, \ldots, \tilde{\boldsymbol{\Phi}}_n]^T$ is a $n \times n$ matrix given $\tilde{\boldsymbol{\Phi}}_i = [\tilde{\mathbf{z}}_i \cdot \tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_i \cdot \tilde{\mathbf{z}}_n]^T$. We can rearrange (14) to be [9]

$$(\mathbf{y} - \tilde{\mathbf{Z}}\tilde{\mathbf{w}})^T (\mathbf{y} - \tilde{\mathbf{Z}}\tilde{\mathbf{w}}) = (\mathbf{y} - \tilde{\mathbf{Z}}\hat{\mathbf{w}})^T (\mathbf{y} - \tilde{\mathbf{Z}}\hat{\mathbf{w}}) + [\tilde{\mathbf{Z}}(\tilde{\mathbf{w}} - \hat{\mathbf{w}})]^T [\tilde{\mathbf{Z}}(\tilde{\mathbf{w}} - \hat{\mathbf{w}})] \tag{15}$$

After careful derivation and consideration of (10)(15) and the assumption in RVM, we write the likelihood function in a form of

$$p(\mathbf{y}|\tilde{\mathbf{w}}, \beta, \gamma) = \frac{1}{(2\pi)^{n/2}} \left\{ \exp\left[-\frac{\beta}{2} (\tilde{\mathbf{w}} - \hat{\mathbf{w}})^T \mathbf{1}_{(d+2)\times n} \mathbf{R}^{-1} \times \mathbf{1}_{n\times(d+2)} (\tilde{\mathbf{w}} - \hat{\mathbf{w}}) \right] \right\} \cdot \left\{ \beta^{\gamma - 1} \exp\left[-\frac{\beta}{\eta} \right] \right\} \tag{16}$$

where $\mathbf{R} = \tilde{\boldsymbol{\Phi}}^{-1}$, $\gamma = (n+2)/2$, $\eta^{-1} = (\mathbf{y} - \tilde{\mathbf{Z}}\hat{\mathbf{w}})^T (\mathbf{y} - \tilde{\mathbf{Z}}\hat{\mathbf{w}})$ and $\mathbf{1}_{ij} = 1$ for all i, j. In (16), the first term is viewed as a Gaussian density for $\tilde{\mathbf{w}}$ given β, and the second term is viewed as a gamma density for β. This distribution is known as normal-gamma distribution.

3.2. Recursive Bayesian for Incremental Learning

To activate incremental learning, we select the prior density of regression parameters $\tilde{\mathbf{w}}$ and β as a *conjugate prior* [3], which is a normal-gamma distribution

$$p(\tilde{\mathbf{w}}, \beta) \propto \text{Normal} - \text{Gamma}(\hat{\mathbf{w}}, \mathbf{R}, \gamma, \eta) . \tag{17}$$

Advantage of considering this conjugate prior is twofold. First, the prior density and the pooled posterior density belong to the same distribution family so that the incremental learning mechanism can be activated. Second, the computation of model parameters is efficient because the mode of posterior density turns out to be MAP estimate. In incremental learning, we collect a sequence of block data $D^T = \{D^{(1)}, \cdots, D^{(T)}\}$ for model adaptation in individual epoch. At learning epoch t, we have current block data $D^{(t)} = \{(\mathbf{z}_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t}$ and sufficient statistics accumulated from previous data sequence $D^{t-1} = \{D^{(1)}, \cdots, D^{(t-1)}\}$. We are calculating the posterior distribution, which combines the likelihood function of current block data $D^{(t)}$ and the prior density calculated from historical data D^{t-1}

$$\begin{aligned}
&p(\tilde{\mathbf{w}}^t, \beta | D^t) = p(\tilde{\mathbf{w}}^{(t)}, \tilde{\mathbf{w}}^{t-1}, \beta | \mathbf{y}^{(t)}, \mathbf{y}^{t-1}) \\
&\propto p(\mathbf{y}^{(t)} | \tilde{\mathbf{w}}^{(t)}, \beta) p(\mathbf{y}^{t-1} | \tilde{\mathbf{w}}^{t-1}, \beta) p(\tilde{\mathbf{w}}^{(t)}, \tilde{\mathbf{w}}^{t-1} | \beta) p(\beta) \\
&= \left\{ \exp\left[-\frac{\beta}{2} (\tilde{\mathbf{w}}^{(t)} - \hat{\mathbf{w}}^{(t)})^T \mathbf{1}_{(d+2)\times n_t} \mathbf{R}^{(t)^{-1}} \mathbf{1}_{n_t \times (d+2)} (\tilde{\mathbf{w}}^{(t)} - \hat{\mathbf{w}}^{(t)}) \right] \right\} \\
&\times \left\{ \exp\left[-\frac{\beta}{2} (\tilde{\mathbf{w}}^{t-1} - \hat{\mathbf{w}}^{t-1})^T \mathbf{1}_{(d+2)\times n^{t-1}} (\mathbf{R}^{t-1})^{-1} \mathbf{1}_{n^{t-1}\times(d+2)} (\tilde{\mathbf{w}}^{t-1} - \hat{\mathbf{w}}^{t-1}) \right] \right\} \\
&\times \left\{ \exp\left[-\frac{\beta}{2} \begin{bmatrix} \tilde{\mathbf{w}}^{t-1} - \hat{\mathbf{w}}^{t-1} \\ \tilde{\mathbf{w}}^{(t)} - \hat{\mathbf{w}}^{(t)} \end{bmatrix}^T \mathbf{1}_{(2d+4)\times n^t} \begin{bmatrix} (\mathbf{R}^{t-1})^{-1} & \mathbf{V} \\ \mathbf{V}^T & (\mathbf{R}^{(t)})^{-1} \end{bmatrix} \right. \right. \\
&\left. \left. \times \mathbf{1}_{n^t \times (2d+4)} \begin{bmatrix} \tilde{\mathbf{w}}^{t-1} - \hat{\mathbf{w}}^{t-1} \\ \tilde{\mathbf{w}}^{(t)} - \hat{\mathbf{w}}^{(t)} \end{bmatrix} \right] \right\} \\
&\times \left\{ \beta^{\gamma^{(t)} - 1} \exp\left[-\frac{\beta}{\eta^{(t)}} \right] \right\} \cdot \left\{ \beta^{\gamma^{t-1} - 1} \exp\left[-\frac{\beta}{\eta^{t-1}} \right] \right\}
\end{aligned} \tag{18}$$

where $\mathbf{V} = \tilde{\mathbf{Z}}^{t-1} \cdot \tilde{\mathbf{Z}}^{(t)}$ and $n^t = n^{t-1} + n_t$. Hence, we have a posterior distribution such as this form

$$p(\tilde{\mathbf{w}}^t, \beta | D^t) \propto \text{Normal} - \text{Gamma}(\hat{\mathbf{w}}^t, \mathbf{R}^t, \gamma^t, \eta^t) , \tag{19}$$

where

$$\hat{\mathbf{w}}^t = \begin{bmatrix} \hat{\mathbf{w}}^{t-1} \\ \hat{\mathbf{w}}^{(t)} \end{bmatrix}, \quad \mathbf{R}^t = \begin{bmatrix} 2(\mathbf{R}^{t-1})^{-1} & \mathbf{V} \\ \mathbf{V}^T & 2(\mathbf{R}^{(t)})^{-1} \end{bmatrix}^{-1}, \quad \eta^t = \frac{\eta^{t-1}\eta^{(t)}}{\eta^{t-1} + \eta^{(t)}}$$

and $\gamma^t = \gamma^{t-1} + \gamma^{(t)}$.

Attractively, the posterior distribution in (19) comes up with a normal-gamma distribution which is the same as the prior distribution determined from previous data D^{t-1}. It is important that the hyperparameters of new normal-gamma distribution provide the updating mechanism of sufficient statistics from $(\hat{\mathbf{w}}^{t-1}, \mathbf{R}^{t-1}, \gamma^{t-1}, \eta^{t-1})$ to $(\hat{\mathbf{w}}^t, \mathbf{R}^t, \gamma^t, \eta^t)$. Using this reproducible normal-gamma distribution, we are able to learn kernel regression parameters for any time when a block of adaptation data is enrolled. Different from SVR and RVM, we don't need to apply evidence framework to find hyperparameters. The existing models can be adapted to the nonstationary environments. Robustness of pattern recognition is guaranteed.

4. EXPERIMENTS

4.1. Experimental Conditions

In this paper, we evaluate the performance of recursive Bayesian regression for face recognition on public domain FERET database. We selected 200 persons from **b** set of FERET database. Each person had 11 face images. Resolution of each image was in 256×384 pixels. Face region was in 138×156 pixels. As shown in Figure 1, **ba**, **bj** and **bk** indicate the frontal images. **bj** and **bk** has different

facial expression and lighting condition, respectively. **bb** through **bi** is a series of images under different pose angles.

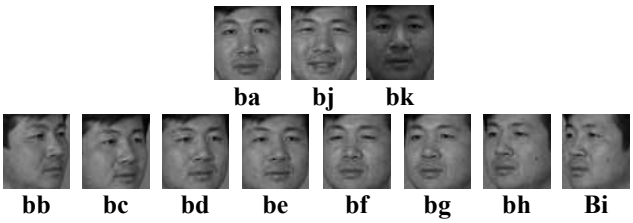

Figure 1: Examples of one person in FERET database.

To test system robustness, for each person, we adopted three images, **ba**, **be**, and **bf,** to train the initial prior model in (16)(17). Number of adaptation data for initial prior model was equivalent to 0. We performed five-fold cross validation through random selection from the other images. For each person, four images were selected for examination of batch learning as well as incremental learning. The remaining four images served as test images. Here, we used RBF kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\sigma_{\text{RBF}}^{-2} \cdot \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right\}. \quad (20)$$

with the given parameter $\sigma_{\text{RBF}}^2 = 100$. The regularization parameter C was obtained by Bayesian SVM method [5].

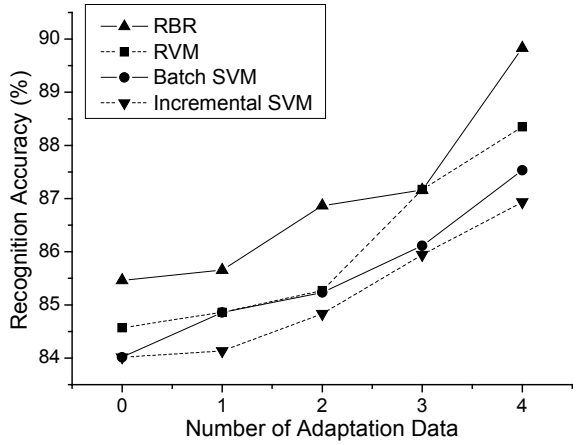

Figure 2: Comparison of recognition accuracies (%) using SVM, RVM and RBR.

4.2. Experimental Results

In Figure 2, we compare the results of the proposed recursive Bayesian regression (RBR) algorithm, SVM and RVM. These methods are based on binary classification, one-against-one strategy. Class number is 200. Batch SVM [13], incremental SVM [4] and RVM [12] were carried out. The proposed RBR method achieves over 85% recognition accuracy. The accuracies of RBR method are better than batch SVM, incremental SVM and RVM. We do see the consistent improvement by increasing number of adaptation data. These results show a significant performance by adapting the facial kernel regression when testing on the data in different environments.

5. CONCLUSION

We have presented a novel recursive Bayesian regression approach for online pattern recognition. Importantly, the relations to SVM, SVR and RVM were illustrated. The updating mechanism of sufficient statistics was developed. The advantages of proposed RVR method were its fast and robust capabilities. We have shown the improvement of face recognition compared to SVM and RVM. The incremental learning improvement from adaptation data was confirmed. The spirit of SVM with maximum margin and minimum error will be illustrated in the future.

6. REFERENCES

[1] J. Bi and K. P. Bennett, "A geometric approach to support vector regression", *Neurocomputing*, vol. 55, nos. 1-2, pp. 79-108, 2003.

[2] J.-T. Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov models", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 268-278, 2002.

[3] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, 1970.

[4] C. P. Diehl and G. Cauwenberghs, "SVM incremental learning, adaptation and optimization", in *Proc. IEEE Int. Joint Conf. on Neural Networks*, vol. 4, pp. 2685-2690, 2003.

[5] J. T.-Y. Kwok, "The evidence framework applied to support vector machines", *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp.1162-1173, 2000.

[6] M. H. Law, J. T. Kwok, "Bayesian support vector regression", in *Proc. 8th Int. Workshop on Artificial Intelligence and Statistics*, pp.239-244, 2001.

[7] Z. Li and X. Tang, "Bayesian face recognition using support vector machine and face clustering", in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 374-380, 2004.

[8] D. J. C. MacKay, "Bayesian interpolation", *Neural Computation*, vol. 4, no. 3, pp. 415-447, 1992.

[9] J. O. Rawlings, S. G. Pantula and D. A. Dickey, *Applied Regression Analysis – A Research Tool*, Springer-Verlag, Inc., 1998.

[10] K. Shima, M. Todoriki and A. Suzuki, "SVM-based feature selection of latent semantic feature", *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1051-1057, 2004.

[11] N. D. Smith and M. J. F. Gales, "Using SVMs and discriminative models for speech recognition", in *IEEE Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 77-80, 2002.

[12] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine", *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001.

[13] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998.