社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

多クラスサポートベクトルマシンの確率的評価値出力手法

リュウリュウ[†]　マウリシオ・ククレ[†]　黒柳　奨[†]　岩田　彰[†]

[†] 名古屋工業大学大学院　工学研究科情報工学専攻　〒466-8555 名古屋市御器所町
E-mail: †liuliu@mars.elcom.nitech.ac.jp, ††mauricio@kugler.com, †††{bw,iwata}@nitech.ac.jp

あらまし　多クラスサポートベクトルマシン (SVM) は認識性能の優れた学習モデルの一つであり，近年多く実用的な認識問題に適用されている。しかし、SVM の出力は正規化されていないため、多クラス SVM においては各 SVM の出力値を比較利用することが困難である。これに対して Platt によって SVM の出力を事後確率に変換することで正規化する手法、シグモイドフィッティングが提案されている。本論文では、多クラス SVM での使用を考慮して Platt の手法を拡張し、全ての SVM のパラメータ最適化を一括して行う手法を提案し、計算機シミュレーションにより本手法の有効性を確認した。

キーワード　サポートベクトルマシン, 多クラス分類問題, 事後確率, シグモイドフィッティング, パラメータ最適化

Probabilistic Outputs for Multiclass Support Vector Machines

Liu LIU[†], Mauricio KUGLER[†],

Susumu KUROYANAGI[†], and Akira IWATA[†]

[†] The authors are with the Department of Computer Science & Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan.
E-mail: †liuliu@mars.elcom.nitech.ac.jp, ††mauricio@kugler.com, †††{bw,iwata}@nitech.ac.jp

Abstract　Support Vector Machines (SVM) have been successfully applied in many classification tasks with great generalization performance. However, the output function of SVMs gives an uncalibrated value, impairing the post-processing and making the combination of several classifiers inefficient, as in the case of multiclass SVMs. Some methods of transforming the binary SVM output in a calibrated posterior probability have been proposed, notably the sigmoid fitting method by Platt. This paper proposes an extension of the Platt's model for multiclass SVMs, by combining the optimization procedures of all sigmoid functions. Experimental results are presented and confirm the efficiency of the proposed method.

Key words　support vector machines, multiclass classification, posterior probability, sigmoid fitting, optimization procedure

1. Introduction

Support Vector Machine (SVM) [1], [2] is a structural risk minimization method that has been successfully applied in many classification tasks with great generalization performance. In order to be applied to multiclass problems, an ensemble of classifiers is necessary, as SVM is an strictly binary classifier. Due to its uncalibrated output function and different output ranges among classifiers, the direct combination of several SVMs in an ensemble structure is inefficient [3]. Moreover, as mentioned by many authors, a classifier should output posterior class probabilities to allow post processing [3], [4]. Many approaches address the problem of converting the SVM output in a calibrated probability, being Platt's methodology [4] the most well know.

Platt's approach consists of the direct conversion of the SVM output function values to posterior probabilities by fitting the SVM output with a sigmoidal function. This solution has the desirable property of maintaining the sparseness of the solution. However, this method is only suitable for binary classifiers, as it is based on a binomial maximum likelihood estimation to fit the sigmoid functions. A natural approach for extending this approach for multiclass domain is fit the output of each binary SVM with an independent sigmoid function, a strategy adopted by several authors [3], [5].

Although giving acceptable results in simple problems, this approach does not guarantee a global optimal solution. Even though each SVM will be fitted by its optimal sigmoidal function, not necessarily these fittings will be the best ones when all the outputs are combined in order to generate the final output. In order to find the best solution, the fitting functions should be optimized in relation to the decoded final output.

Another problem with this naive approach happens when some categories presents small number of samples (unbalanced data) or presents a high separability. The first case lead to biased sigmoidal functions, while, for the second, any parameters gives perfect separations, making the optimization method to diverge or give extreme values. Platt recommended the use of a cross-validation procedure in order to avoid this problems of convergence. Another solution is the use of extra noisy data. Those procedures. however, do not solve the problem completely, also increasing the computational complexity of the classifier training.

This work proposes a new method for optimizing the sigmoidal functions. Instead of independently fitting each SVM's output, the proposed approach combine all sigmoidal function parameters in a single optimization procedure. Thus, all sigmoidal functions' parameters are retrieved simultaneously. Moreover, the interdependency of the functions avoid the optimization diverge for unbalanced or too simple classifiers.

Experimental results show that for balanced problems, the proposed method presented comparable results to the naive application of independent sigmoidal functions. For unbalanced classes, the new approach outperformed the previous approach considerably.

The organization of the paper goes as follows: a more detailed revision of Platt's model and previous extensions to multiclass are presented in Section 2., and Section 3. introduces the proposed model. Section 4. presents experiments with the new model and comparisons with previous methods, and Section 5. concludes the paper with analysis of the results and suggests possible future extensions.

2. Fitting the SVM outputs

2.1 Original Platt's model

After the SVM training procedure, Platt suggested to use a two-parameter sigmoid function in order to change the SVM output $f(\mathbf{x})$ in a posterior probability $P(\omega_k \,|\mathbf{x})$. The use of a sigmoidal function comes from empirical observations of the distribution of the output values. This approach keeps the SVM error function unchanged, also maintaining the sparseness of the solution. The sigmoidal function has the form:

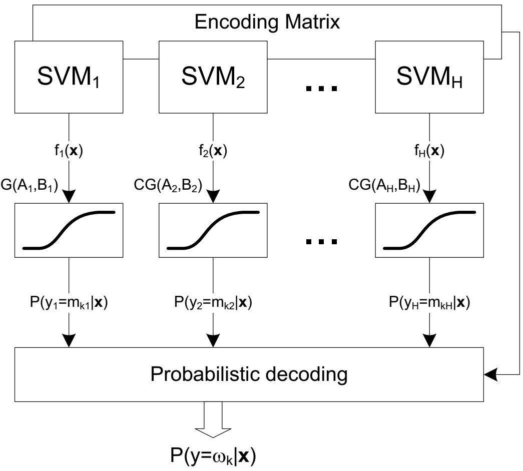

図 1 Independent sigmoid fitting multiclass SVM structure

$$P(y=1\,|f\,) = \frac{1}{1+\exp(Af+B)} \tag{1}$$

in which f is a simplified notation of the SVM output given by:

$$f(\mathbf{x}) = \sum_{n\in SV} y_n \alpha_n K(\mathbf{x}_n, \mathbf{x}) + b \tag{2}$$

where $\mathbf{x}_n$ is the n^{th} support vector, y_n is the label of the n^{th} support vector, $K(\mathbf{x}_n, \mathbf{x})$ is the Kernel function, α_n is the Lagrange multiplier of the n^{th} support vector and b is the bias.

In order to find an optimal sigmoid fitting based on the input patterns, the posterior class probability $p(y\,|f\,)$ is maximized by the *likelihood function* of the SVM output:

$$L'(y_i\,|p_i\,) = \prod_i P_i(Y_i = y_i\,|p_i\,) \tag{3}$$

where Y_i is the classifier's final output and p_i is an abbreviated form of equation 1. Redefining y_i in order to represent a target probability:

$$t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & if \quad y_i = +1 \\ \frac{1}{N_- + 2} & if \quad y_i = -1 \end{cases} \tag{4}$$

where N_+ is the number of positive samples and N_- is the number of the negative samples. Finally, with the use of the new targets, equation 4 assumes the form of a Bernoulli distribution and can be rewritten as:

$$L'(t_i\,|p_i\,) = \prod_i p_i^{t_i} (1-p_i)^{1-t_i} \tag{5}$$

or, in the negative log likelihood form (abbreviating $-\ln L'(t_i\,|p_i\,)$ to L):

$$L = -\sum_i \left[t_i \ln(p_i) + (1-t_i)\ln(1-p_i)\right] \tag{6}$$

The parameters $\hat{A}$ and $\hat{B}$ that minimize equation (6) are

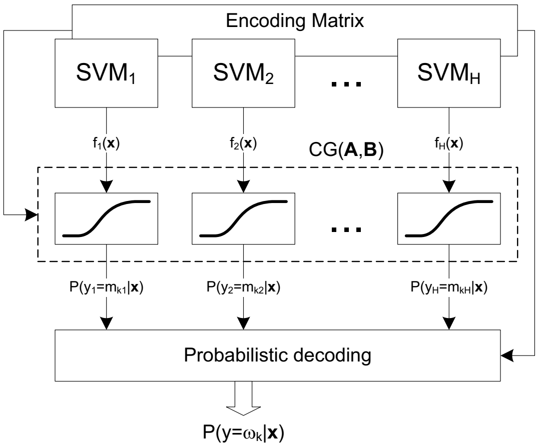

図 2 Multiclass sigmoid fitting multiclass SVM structure

the best fitting sigmoid function for $f(\mathbf{x})$. In order to perform this minimization, Platt used a model-trust minimization algorithm in his experiments. In this paper, the Conjugate Gradient (CG) minimization method [6] was used.

2.2 Previous extensions to multiclass classification and decoding frameworks

Several authors proposed extensions of Platt's model for multiclass classification. All these methods, however, naively apply his method in each SVM independently, consisting basically in frameworks for decoding the various binary classifiers probabilities in a final posterior class probability output. For instance, Passerini, Pontil and Frasconi [3] proposed a new probabilistic decoding procedure for multiclass SVM using error correcting output encodings that outperformed other decoding methods, such as hamming distance and loss based decoding. This method, however, considers all classes statistically independent, what is not true, as the classifiers corresponding to one class were trained with the same samples of that class.

In their large scale classification model CombNET-III, Kugler *et al.* [5] proposed an alternative probabilistic decoding function which maintains the classifiers confidence on the overall sample space, a desirable property for divide-and-conquer based models.

Given a coding matrix $\mathbf{M}^{K\times H}$ in which K is the number of classes and H is the number of classifiers, $m_{k,h} = \{-1, 0, +1\}$ and zero entries are interpreted as "don't care", the probability of class ω_k given an unknown sample $\mathbf{x}$ is defined as the average probability outputted by the classifiers containing that class. The proposed decoding function hence becomes:

$$P(\omega_k|\mathbf{x}) = \frac{\sum\limits_{h:m_{k,h}\neq 0} P(y_h = m_{k,h}|\mathbf{x})}{\sum\limits_{h=1}^{H} |m_{k,h}|} \tag{7}$$

where $\mathbf{M}^{K\times H}$ is the coding matrix, with $m_{k,h} = \{-1, 0, +1\}$, K is the number of classes and H is the number of classifiers.

This model, however, also presents all the SVMs sigmoid functions being optimized independently. The main structure of this kind of model is shown in Figure 1. All proposed decoding functions rely on well estimated sigmoid functions. However, as mentioned before, situations in which one class has considerably more samples than the other or two classes can be separated much more easily than others are hard to deal and often leads to bad classification results. Next section introduces a method for addressing all this problems.

3. Proposed Model

In order to combine the optimization of all sigmoid functions in a single process, a new likelihood function must be defined. The likelihood function must be based on a probabilistic decoding framework that combines all SVM output probabilities in a final output. This paper makes use of the decoding function proposed by Kugler *et al.* [5] and shown in equation (7), although any other decoding function could be used, following the same procedure.

At first, the decoding function must be modified in order to directly contain the two-parameter sigmoid function of equation (1). Considering the two possible values of m_{kh} on the upper part of equation (7), it is possible to denote:

$$P(y_h = m_{k,h}|\mathbf{x}_i) = \begin{cases} p_{ih} & m_{kh} = +1 \\ 1 - p_{ih} & m_{kh} = -1 \end{cases} \tag{8}$$

The two situations can be combined as:

$$P(y_h = m_{k,h}|\mathbf{x}_i) = \frac{1 - m_{kh}}{2} + m_{kh}p_{ih} \tag{9}$$

Applying (9) in (7):

$$P(\omega_k|\mathbf{x}_i) = \frac{\sum\limits_{h:m_{k,h}\neq 0} \frac{1-m_{kh}}{2} + m_{kh}p_{ih}}{\sum\limits_{h=1}^{H} |m_{k,h}|} \tag{10}$$

where

$$\begin{aligned} p_{ih} &= P_{A,B}(y = 1|f_h(\mathbf{x}_i)) \\ &= \frac{1}{1 + \exp(A \cdot f_h(\mathbf{x}_i) + B)} \end{aligned} \tag{11}$$

The new target probabilities must now be defined for all classes. Following the same approach proposed by Platt for defining the values of the targets based on the number of samples:

$$t_{ik} = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & if \quad \mathbf{x}_i \in \omega_k \\ \frac{1}{N_- + 2} & if \quad \mathbf{x}_i \notin \omega_k \end{cases} \tag{12}$$

where the value of each element of the target probability vector tend to $\{0, 1\}$ when the number of samples tends to infinite and to 0.5 when the number of samples tend to zero. As

表 1　Database Description

Database	Classes	Training	Test	Features
Segment	7	210	2100	18
Optdigits	10	3823	1797	64
Satimage	6	4435	2000	36

the targets are now represented by a vector with k elements, it is not possible to use a binomial distribution. Instead, the proposed likelihood function is defined as:

$$L'\left(\mathbf{t}_i\left|\mathbf{p}_i\right.\right) = \prod_i \prod_k \left[P\left(\omega_k\left|\mathbf{x}_i\right.\right)\right]^{t_{ik}} \tag{13}$$

or, in the negative log likelihood form (abbreviating $L'\left(\mathbf{t}_i\left|\mathbf{p}_i\right.\right)$ to L):

$$L = -\sum_i \sum_k \left[t_{ik} \ln \frac{\sum\limits_{h:m_{k,h}\neq 0} \frac{1-m_{kh}}{2} + m_{kh}p_{ih}}{\sum\limits_{h=1}^{H} \left|m_{k,h}\right|} \right] \tag{14}$$

By finding the first and second derivatives of equation (14), it is possible to apply the CG minimization method in order to find the optimal parameters $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. The final structure of the proposed method is shown diagrammatically in Figure 2.

4. Experiments

Three databases from the UCI repository [7] were used to evaluate the performance of the proposed method. The *Segment* database is an image segmentation database containing 3x3 pixel regions extracted from 7 outdoor images (categories)[注1]. The *Optdigits* is a handwritten digits recognition database. The *Satimage* database consists of the multi-spectral values of pixels in 3x3 neighborhoods in satellite images of several types of soil[注2]. A description of the used databases is shown in Table 1.

All experiments were performed using in-house developed software packages. The output encoding used on the experiments is the One-versus-One (OvO) [8], although any other encoding could have been used, as the proposed method is generic for any encoding matrix. The SVM used Kernel function was the Gaussian Kernel. The proposed method was compared with the the results obtained by simply fitting an independent sigmoid in each SVM and using equation (7) to decode the final probability.

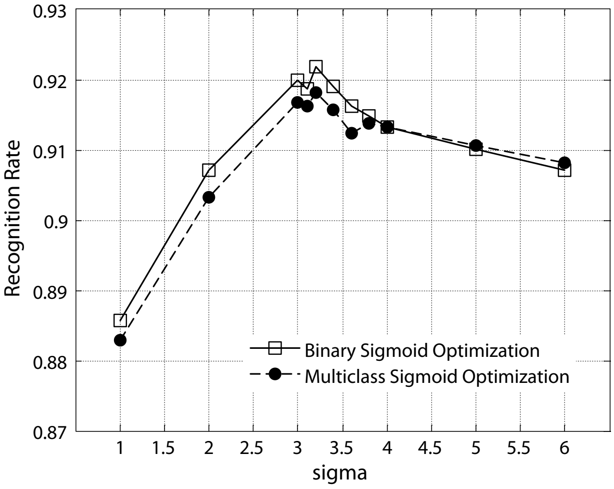

図 3　Classification results for the *Segment* database

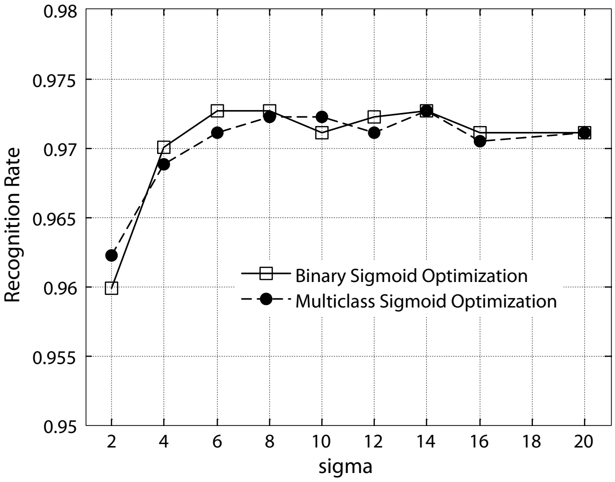

図 4　Classification results for the *Optdigits* database

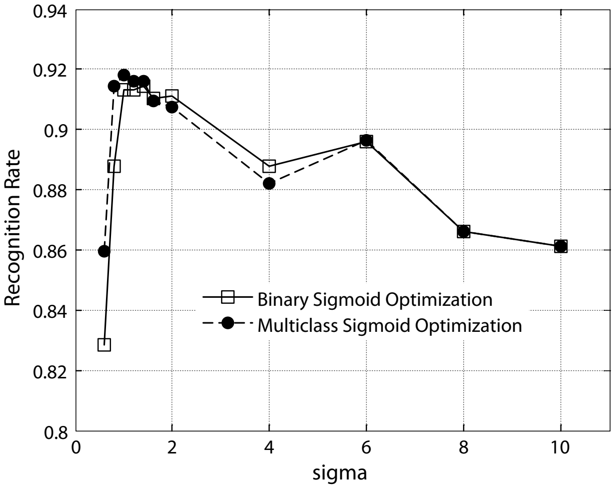

図 5　Classification results for the *Satimage* database

Figure 3 shows the classification accuracy obtained by both methods for the *Segment* database for several values of the Kernel parameter σ. The value of the soft-margin parameter was experimentally fixed in $C = 20$ for both methods. The proposed method maximal accuracy was 91.81% while the independent binary fitting method's maximal accuracy was

[注1] : For the *Segment* database, the feature that contains the number of pixels (”region-pixel-count”) is constant (always 9) and was removed, changing the original number of features from 19 to 18.

[注2] : The *Satimage* database have one of the 7 classes with no patterns. This class was not considered and the classes were relabeled from 0 to 5.

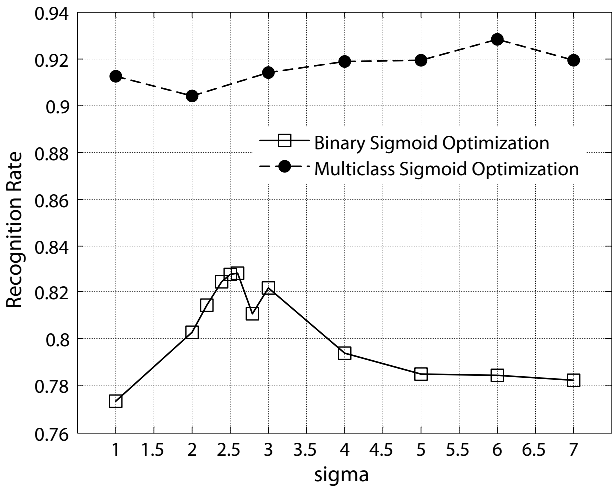

図 6 Classification results for the unbalanced *Segment* database (with a single training sample in the last category)

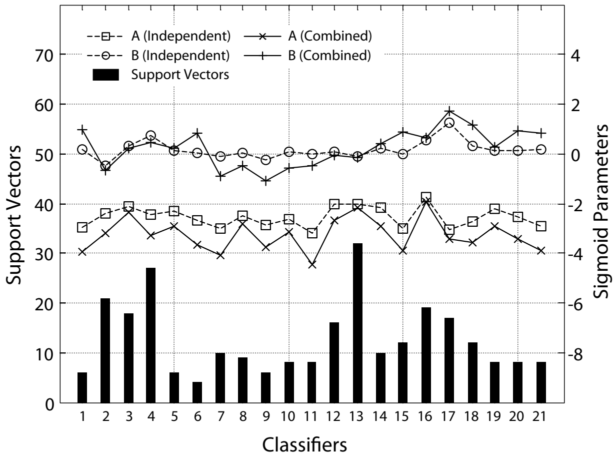

図 7 Sigmoid fitting parameters for the *Segment* database

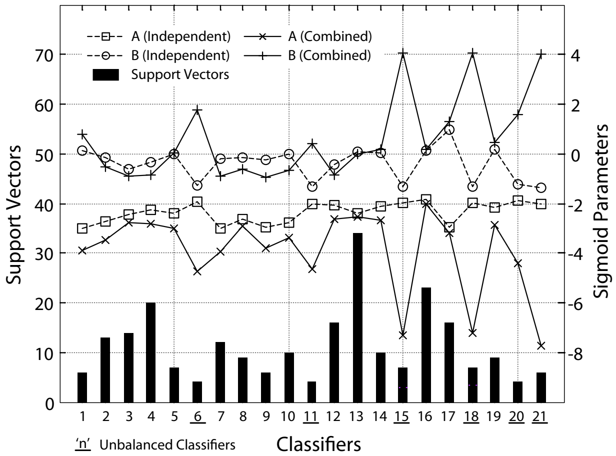

図 8 Sigmoid fitting parameters for the unbalanced *Segment* database (with a single training sample in the last category)

92.19%. Figure 4 shows the classification accuracy obtained by both methods for the *Optdigits* database. Again, the soft-margin parameter was experimentally fixed in $C = 60$. The maximal accuracies for the proposed and the independent fitting methods were both equal to 97.27%. Finally, Figure 5 shows the accuracy results for the *Satimage* database. The maximal accuracies for the proposed and the independent fitting methods were respectively 91.80% and 91.45%. As it can be observed, there is no statistically relevant difference on the accuracy obtained by both methods.

In order to analyze the behavior of the proposed method for unbalanced data, the *Segment* database was modified, leaving a single training sample on the last class "Grass", resulting in a proportion of 30:1 of unbalance between this class and the other 5 classes, each with 30 training samples. The test data was kept complete. Figure 6 show the results of this experiment. The independent fitting method's accuracy dropped almost 10%, while the proposed method's accuracy presented no significant variation.

Figures 7 and 8 shows the analysis of the sigmoidal functions parameters for each of the classifiers, respectively for the original segment data and the unbalanced one. On the original database, all classifiers are trained with 60 samples (30 for each class). On the unbalanced case, as the last class has just a single sample, some classifier (which have their labels underlined on the x-axis on Figure 8) have 31 training samples. The small differences on the number of support vectors and sigmoidal parameters for the unchanged classifiers are due the different average and standard deviation found on the normalization procedure.

In Figure 7, both methods generate similar sigmoidal functions. For the unbalanced case, however, the independent fitting method presents only small change for both parameters, while the proposed method presents significantly higher values of B and more negative values of A for the classifier with large difference between both classes' training samples number. This, together with the higher accuracy obtained confirms the ability of the proposed method on adapt the sigmoid functions in order to maximize the global generalization.

5. Discussion and Conclusions

This paper proposed a new method for obtaining probabilistic outputs for multiclass SVM. The method, based on Platt's method of fitting a sigmoidal function on the output of a binary SVM, differ from traditional approaches that naively fit each SVM independently. Instead, a single optimization procedure maximizes the decoded posterior probabilities.

Several experiments showed that, for simple databases, the method presents comparable results with the naive approach, with no statistically significant difference. For the unbalanced problem used on the experiments, the proposed method outperformed the previous method, presenting a higher accuracy. Analysis of the obtained sigmoidal function parameters shows that the new method successfully adapted the fitting functions on the unbalanced classifiers.

Future works include the application of the proposed model to the large scale classifier CombNET-III, as it often presents problems of unbalanced categories on its multiclass SVM based branch network. Also, a computational complexity analysis of the new method is still necessary.

文献

[1] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol.20, no.3, pp.273–297, 1995.

[2] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press, Cambridge, 2000.

[3] A. Passerini, M. Pontil, and P. Frasconi, "New results on error correcting output codes of kernel machines," IEEE Transactions on Neural Networks, vol.15, no.1, pp.45–54, January 2004.

[4] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, ed. A.J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, pp.61–74, MIT Press, Cambridge, MA, March 1999.

[5] M. Kugler, S. Kuroyanagi, A.S. Nugroho, and A. Iwata, "CombNET-III: a support vector machine based large scale classifier with probabilistic framework," IEICE Transactions on Information & Systems, vol.E89-D, no.9, pp.2533–2541, September 2006.

[6] J.R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Available at http://www-2.cs.cmu.edu/∼quake-papers/painless-conjugate-gradient.pdf, August 1994.

[7] C.L. Blake and C.J. Merz, "UCI repository of machine learning databases." Irvine, CA: University of California, Department of Information and Computer Science, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[8] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," Advances in Neural Information Processing Systems, ed. M.I. Jordan, M.J. Kearns, and S.A. Solla, The MIT Press, 1998.