Feature Subset Selection for Support Vector Machines using Confident Margin

Mauricio Kugler, Kazuma Aoki, Susumu Kuroyanagi, Akira Iwata
Grad. School of Engineering, Dept. of Computer Science & Engineering
Nagoya Institute of Technology
Gokiso-cho, Showa-ku Nagoya, Japan, 466-8555
E-mail: mauricio@kugler.com, bw@nitech.ac.jp

Anto Satriyo Nugroho
School of Life System Science & Technology
Chukyo University
101 Tokodachi Kaizu-cho Toyota, Japan, 470-0393
E-mail: nugroho@life.chukyo-u.ac.jp

***Abstract*— The aim of this study is to develop a feature subset selection (FSS) method based on the margin of Support Vector Machines (SVM). The problem of directly using the SVM margin is that it does not always provide clear relationship between its value and the performance of SVM, and the best obtained subset is not guaranteed to be the best possible one. In this paper, a new solution is describe by the introduction of the *Confident Margin* (CM) in the subset criterion, which permits to get near the best recognition rate by monitoring the peak of CM curve without directly calculating the recognition rate, in order to save computational time. The performance of the proposed method was evaluated in artificial and real-world data experiments.**

I. Introduction

The application of pattern recognition in the real-world domain data often encounters problems caused by the high dimensionality of the input space. The situation becomes worse if the ratio of relevant features to the irrelevant ones is low. By removing these insignificant features, the learning process becomes more effective, and the performance of the classifier will be increased. This is the motivation of using feature subset selection in a pattern recognition system.

Feature subset selection (FSS) refers to algorithms that select the most relevant features to the classification task, removing the irrelevant ones. Two aspects are important in designing a feature subset selection: selection algorithm and selection criterion. Based on the selection algorithm, various methods have been proposed. Jain [1] provided a useful taxonomy of selection algorithms. The present study is limited to the methods belonging to the sequential selection algorithms, which will be applied for classification task using Support Vector Machines.

Support Vector Machines (SVM) have aroused many attentions in pattern recognition field. Many studies reported it superiority to the conventional methods. The application of SVM is also found in a broad spectrum of technology, including data mining and bioinformatics [2][3][4]. In the case of its application to the real-world domain data, often provided as large scale dataset, the computational cost of SVM is expensive. This issue should be carefully considered in developing the feature subset selection for classification using SVM.

In this context, it was developed a feature selection algorithm which is based on particular characteristics of the SVM. A new feature selection criterion function was proposed, based on the SVM's margin, in the hope that it provides a better representation of the expected SVM's performance. The performance of the proposed method was verified through several experiments, and it was also compared with the Recursive Feature Elimination [2] and a leave-one-out recognition rate based method.

II. Feature Subset Selection using Confident Margin

Let Y be the set that comprises all the n original features, and $X(X \subseteq Y)$ a selected subset of Y, with d features. $J(X)$ is a function that determines how good the subset X is, by a certain criterion.

The problem of FSS is defined by searching the subset $X \subseteq Y$ composed by d features ($d = |X|$), that satisfies:

$$J(X) = \max_{Z \subseteq Y, |Z|=d} J(Z) \tag{1}$$

Various criterion functions J have been proposed to measure the quality of the subset of the features. Based on the selection criterion, the methods of FSS can be categorized into two approaches: filter and wrapper. The filter methods [5] employ the intrinsic properties of data, such as class separability. The wrapper methods [6] evaluates the quality of the subset based on the classification rate of a classifier, calculated by methods such as the well-known leave-one-out (LOO) or cross validation. Generally, the wrapper method achieves better performance than that of filter method, but the computational cost is expensive, as they require the retraining of the classifier and, in the absence of test data, the evaluation of the correctness rate by computationally expensive procedures.

In this study, an "unsupervised" wrapper method were implemented. Even it still need the retraining of the SVM, it is not necessary to evaluate the correctness rate, as the generalization performance is evaluated by and indirect measurement.

A. Normal Margin

The concept of margin plays an important role in SVM theory. Margin (or sample-margin) measures the distance between the instances and the decision boundary induced by the SVM [5]. The aim of the training phase in SVM is to maximize

the margin, in order to obtain the optimal hyperplane in the feature space. In this paper, the geometric margin of SVM is named as Normal Margin(NM), in order to distinguish it from the proposed criterion.

Let $\mathbf{x}_i \in \mathbb{R}^n$ $(i = 1, \dots \ell)$ be a feature vector, where ℓ is the number of examples, and $y_i \in \{-1, 1\}$ the class assigned to each example. The discrimination function of SVM is given by:

$$f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b \tag{2}$$

The weight vector $\mathbf{w}$ is obtained as follows:

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \Phi(\mathbf{x}_i) \tag{3}$$

where α are the Lagrange multipliers.

Using the weight vector $\mathbf{w}$, the geometrical interpretation of the Normal Margin (NM) is defined as [7]:

$$NM = \frac{1}{\|\mathbf{w}\|} \tag{4}$$

and $\|\mathbf{w}\|$ is given by:

$$\|\mathbf{w}\|^2 = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_j) \tag{5}$$

As $\Phi(\mathbf{x}_1) \Phi(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2)$, by substituting equation (5) in equation (4), it is possible to rewrite the Normal Margin as follows:

$$NM = \left(\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) \right)^{-\frac{1}{2}} \tag{6}$$

In this study, the Gaussian Kernel was used, which is defined as:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2} \right) \tag{7}$$

In the work of Guyon [2], the weight vector norm $\|\mathbf{w}\|$ is used as criterion for subset evaluation as follows:

$$\begin{aligned} &\left| \|\mathbf{w}\|^2 - \left\|\mathbf{w}^{(i)}\right\|^2 \right| \\ &= \frac{1}{2} \left| \sum_{j,k=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_j, \mathbf{x}_k) \right. \\ &\quad \left. - \sum_{j,k=1}^{\ell} \alpha_j^{(i)} \alpha_k^{(i)} y_i y_j K^{(i)}(\mathbf{x}_j, \mathbf{x}_k) \right| \end{aligned} \tag{8}$$

where $K^{(i)}$ and $\alpha_j^{(i)}$ are defined, respectively, as the Kernel function values and the Lagrange multipliers obtained in the absence of the *i*th feature.

When the worst feature *k*th is found, it is removed, which gives a score, named SR (equation (9)), for that subset, calculated from $K^{(k)}$ and $\alpha_j^{(k)}$. This method is called Recursive Feature Elimination (SVM-RFE).

$$SR = \left| \|\mathbf{w}\|^2 - \left\|\mathbf{w}^{(k)}\right\|^2 \right| \tag{9}$$

The use of NM as criterion for subset evaluation is accompanied by the problem that its behavior in the presence of misclassified data. These data will become noisy support vectors that make the NM value not to properly represent the classifier generalization performance. Even the features are ranked, the maximal value of SR does not correspond to the best feature subset, still requiring a monitoring of the classifier accuracy, as it is done in [2] and it is shown in the experiments of section III. This accuracy computing can be computationally expensive if some cross validation method is required (e.g. LOO or *n*-fold cross validation) or the dimensionality is large, requiring longer time in the Kernel function computing.

B. Confident Margin

In order to make the feature selection criterion function proportional to the classifier generalization ability, penalizing the samples proportionally to its function value, a new criterion function, named Confident Margin (CM), will be introduced:

$$CM = c \cdot NM \tag{10}$$

where c is the average confidence of all samples, defined as:

$$c = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i f(\mathbf{x}_i) \tag{11}$$

The distance of a sample $\mathbf{x}$ to the hyperplane is given by:

$$dist = \frac{f(\mathbf{x})}{\|\mathbf{w}\|} \tag{12}$$

Hence, the geometrical interpretation of the CM is that it is the average absolute distance of all samples to the hyperplane.

In this sense, how much closer to the hyperplane or specially how much more misclassified samples one feature subset generates, smaller will be the value of the confidence. Furthermore, similar confidence subsets will be discriminated by the value of the normal margin itself. Hence, it is expected that this new measurement gives better results in the FSS task.

C. Sequential Backward Selection using Confident Margin

Applying the idea of the Confident Margin to a sequential FSS method, a new selection algorithm is proposed, summarized in Figure 1.

For the sake of simplicity, the proposed algorithm will be referred to as SBS-CM (Sequential Backward Selection using Confident Margin). The selection strategy used in this algorithm is developed based on the work of Marill and Green [8], which works in top-down fashion. The selection process starts from a full set of feature, then removes sequentially the most irrelevant ones. To find the most irrelevant feature of

1.Initialize:
Subset of surviving features $\mathbf{s} = [1, 2, \ldots, n]$
2.**repeat**
(a) **for** $\forall s_i \in \mathbf{s}\,(1 \leq i \leq |\mathbf{s}|)$
do train the SVM classifier without ith feature
/*the Kernel matrix is calculated in advance*/
do compute $J_i = CM^{(i)}$
/*$CM^{(i)}$ is the Confident Margin without ith feature.*/
(b) Find the worst feature k
$$k = m \,\Big|\, J_m = \arg\max_q (J_q)$$
(c) Remove the feature that maximizes CM
$\mathbf{s} = [1, \ldots, k-1, k+1, \ldots, n]$
3.**until** $\mathbf{s}$ is empty

Fig. 1. Sequential Backward Selection using Confident Margin (SBS-CM) Algorithm

the current surviving subset, one of the features (e.g. the ith feature) is removed and the CM is calculated. This is denoted as $CM^{(i)}$, i.e. the Confident Margin without ith feature. The ith feature is returned to the subset, and the same procedures are carried out for the other features. Finally, the most irrelevant feature, which its removal produced the greatest value of CM, can be found. The procedure is repeated until all of the features are removed. By monitoring the peak point of CM curve, it is expected to be possible to identify the best subset which has the maximum generalization performance of the generated ranking, without directly calculating the recognition rate.

III. Experimental Results & Discussion

The performance of proposed algorithm were evaluated in several experiments. The first experiment, which used artificial data, helped to analyze how the algorithm works. The later ones were conducted using real-world data obtained from UCI repository [9]. For all the experiments, the used parameters where selected by several tries with the full feature set data.

A. Experiment 1: Artificial Data

In the first experiment, the performance of the proposed method was evaluated in an XOR problem. Two dimensional data were randomly generated uniformly distributed within the intervals as shown in Table I. To the data were added 98 noisy random features (uniformly distributed in the interval $[0, 1]$), making up a 100-dimensional artificial data. The problem is defined by applying feature subset selection to obtain the significant features, i.e. the first two features. Parameters of SVM during the training phase were $\sigma = 1.0$ and $C = 100$.

The result is shown in Figure 2. The horizontal axis shows the number of features (dimensionality of the data), while the vertical one shows the CM value for the current subset. The noise features were removed sequentially, and the two

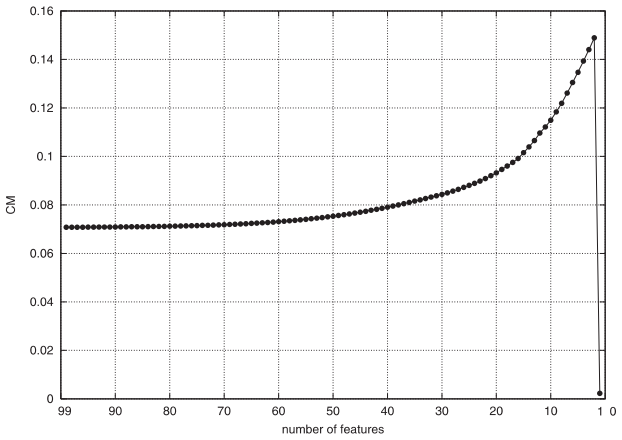

Fig. 2. Confident Margin of SBS-CM in the experiment using the artificial XOR data

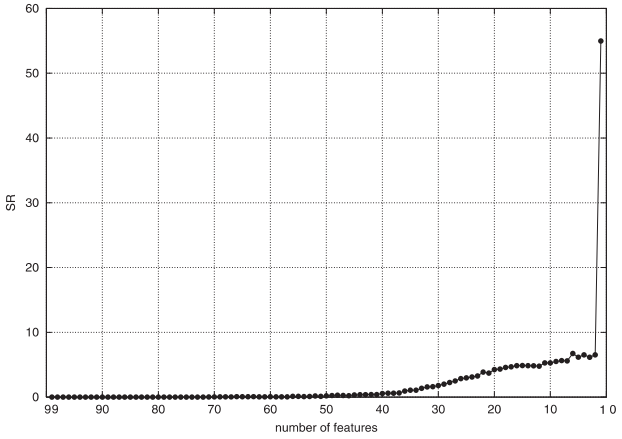

Fig. 3. SR score of SVM-RFE in the experiment using the artificial XOR data

significant features were retained until the final phase of the selection process. When only these two features were remained, the score obtained by the algorithm reached its maximum.

The same experiment was conducted using SVM-RFE and the result is compared with that of the proposed algorithm. Similar to the previous experiment, SVM-RFE also removed the noise features sequentially, and finally two significant features were retained. The score SR of each state during the feature selection process is obtained by equation (8), and depicted in Figure 3. Despite of the features were ranked

TABLE I
Artificial Data

Class	Dim. 1	Dim. 2	Number of Samples
class 1	[0, 0.2]	[0, 0.2]	50
class 1	[0.8, 1.0]	[0.8, 1.0]	50
class 2	[0, 0.2]	[0.8, 1.0]	50
class 2	[0.8, 1.0]	[0, 0.2]	50

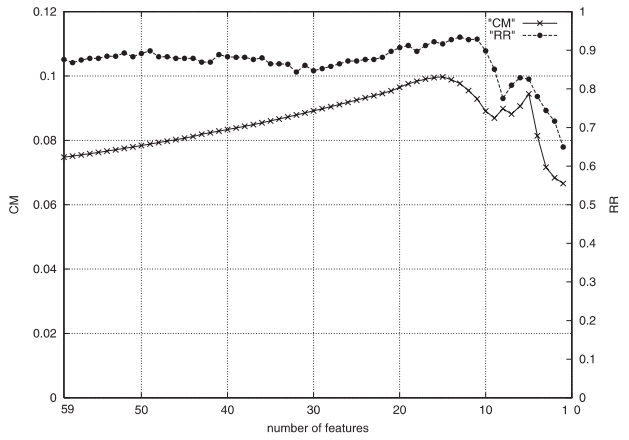

Fig. 4. Confident Margin and Recognition Rate of SBS-CM in the experiment using the *sonar* dataset

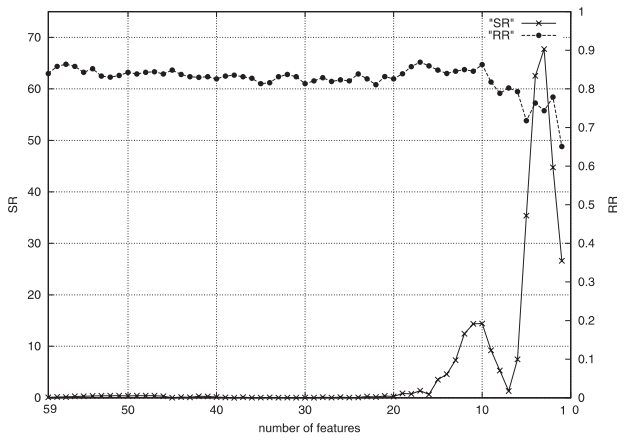

Fig. 5. SR and Recognition Rate of SVM-RFE in the experiment using the *sonar* dataset

correctly, the peak of the curve in Figure 3 does not represent the best feature subset.

In a situation in which the significant features are known in prior (such as Experiment 1), the performance of FSS can be evaluated by confirming if these features are retained in the selected subset. In general, however, it is unknown which of the features has significant information, so, is preferred that these features can be identified by monitoring the score of criterion used in the FSS. The SBS-CM result shows that, in this preliminary experiment, the requirement that the criterion function peak point has to correspond to the best feature subset is fulfilled, while in the case of the SBS-RFE, it is not.

B. Experiment 2 : Real-world Data

Two datasets are chosen from UCI: *sonar* and *ionosphere*. *Sonar* dataset comprises of 208 patterns, 97 of them belonging to class 1 and 111 belonging to class 2. Each of the pattern has 60 features. *Ionosphere* dataset comprises of 351 patterns, 225 of them belonging to class 1 and 126 to class 2. Originally, this database have 34 features, however, since all of the 2nd feature values are zero, this feature were removed. Thus, the samples in *ionosphere* database were represented by 33 features.

Experiment 2a: *Sonar* dataset

Experiment was conducted by applying SBS-CM to the *sonar* dataset, with the SVM parameters $\sigma = 1.8$ and $C = 10$, and the LOO procedure were used to evaluate the recognition rate. The result is depicted in Figure 4. Horizontal axis shows the number of features, the left vertical axis shows the CM value and the right vertical axis shows the Recognition Rate (RR). The same experiment was conducted using SVM-RFE, and the result is shown in Figure 5.

When the CM curve of SBS-CM achieved its peak, the recognition rate is 92% and 15 features were selected. Analyzing the recognition rate curve, the peak is achieved at 93% with 13 features selected. Figure 4 shows that CM and RR graphics have similar behaviors. Consequently, the monitoring of the peak point of Confident Margin curve could had been

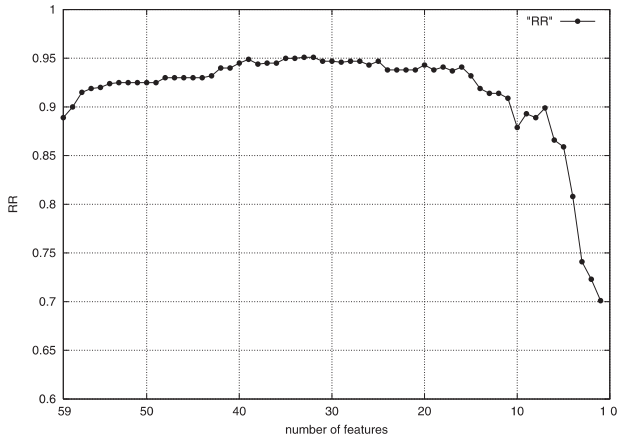

Fig. 6. Leave-one-out Recognition Rate of SBS-LOO in experiment using the *sonar* dataset

used for evaluating the subset which is expected to produce near the best recognition rate of SVM by this ranking criterion.

In the case of the result obtained by SVM-RFE, when SR curve achieves its peak, the recognition rate is 74%, with 3 features. Analyzing the recognition rate curve, its peak is achieved at 87% with 17 features, which is worst than the proposed method result. Figure 5 also shows that the curves of SR and RR do not have similar behavior as in the case of SBS-CM curve. Hence, the monitoring of the SR peak of SVM-RFE does not guarantee that the classifier achieves the

TABLE II
RECOGNITION RATE AT THE PEAK OF CRITERION SCORE AND ITS MAXIMUM OF EACH METHOD IN *sonar* DATASET EXPERIMENT

Feature Selection	Max. criterion value's RR		Best RR	
	#DIM	RR[%]	#DIM	RR[%]
SBS-CM	15	92	13	93
SVM-RFE	3	74	17	87
SBS-LOO	-		32, 33	95

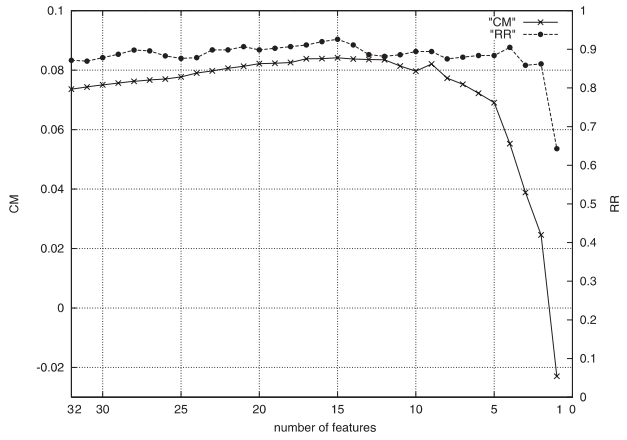

Fig. 7. Confident Margin and Recognition Rate of SBS-CM in the experiment using the *ionosphere* dataset

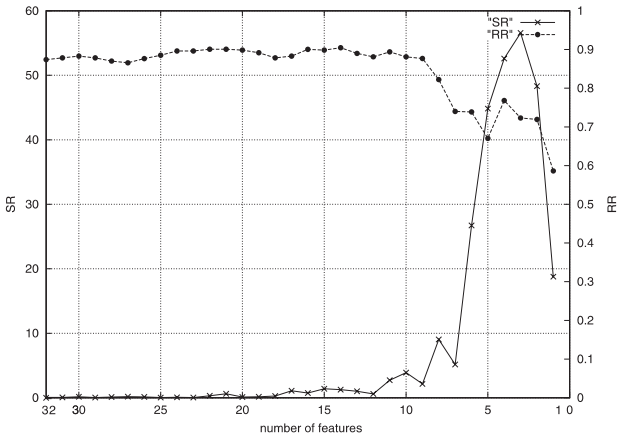

Fig. 8. SR and Recognition Rate of SVM-RFE in the experiment using the *ionosphere* dataset

best recognition rate.

The next experiment was conducted by SBS-LOO, which calculates the recognition rate using LOO for each subset. The SBS-LOO results are shown in Figure 6. The best recognition rate of SBS-LOO is 95%, which is the highest among the three algorithms. It is achieved by reducing the dimensionality up to 32 and 33. However, all the required extra SVM trainings for the LOO procedure make the method very computationally expensive. Table II summarizes the experimental results using *sonar* dataset.

Experiment 2b: *Ionosphere* dataset

The performance of the proposed algorithm was also evaluated using *ionosphere* dataset. The same experiments of the Experiment 2a were conducted by setting $\sigma = 5.0$ and $C = 10$. The result is shown in Figure 7. At the peak point of CM curve, the recognition rate achieved by SBS-CM is 93% with 15 features. The curves of CM and RR show that they achieved their peak at the same number of features.

In the case of SVM-RFE, at the peak point of SR, the recognition rate was 72%, achieved by reducing the dimensionality up to 3 features. Analyzing the RR curve shows that its peak is achieved at 14 features, but the recognition rate, 90%, is smaller than the one achieved by the SBS-CM. Figure 8 also shows that SR and RR curves of SVM-RFE do not have similar behavior.

TABLE III
RECOGNITION RATE AT THE PEAK OF CRITERION SCORE AND ITS MAXIMUM OF EACH METHOD IN *ionosphere* DATASET EXPERIMENT

Feature	Max. criterion value's RR		Best RR	
Selection	#DIM	RR[%]	#DIM	RR[%]
SBS-CM	15	93	15	93
SVM-RFE	3	72	14	90
SBS-LOO	-		23, 24	92

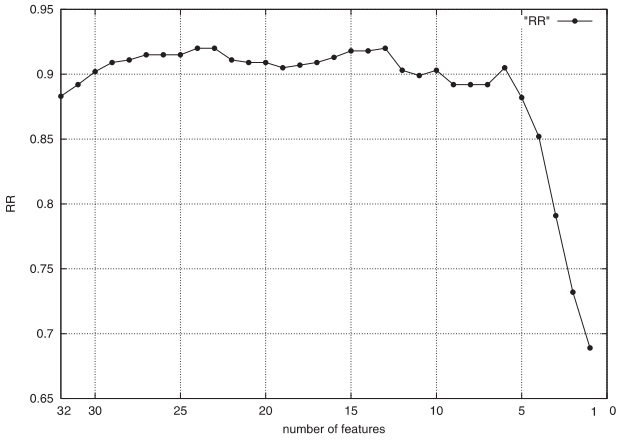

Fig. 9. Leave-one-out Recognition Rate of SBS-LOO in the experiment using *ionosphere* data set

The result of the proposed method was then compared to that of SBS-LOO, depicted in Figure 9. The best recognition rate of this method is 92% for 23 and 24 features. This score is lower than that of SBS-CM. Thus, in this experiment, the proposed method outperforms SBS-LOO in term of recognition rate and dimensionality reduction. Table III summarized the results obtained by these methods in experiment using *ionosphere* dataset.

That obtained results are difficult to compare with previous works that used the same databases for experiments, such as [10] (multilayer perceptron, 10-fold cross-validation; *sonar*: 7 features, 75% recognition rate; *ionosphere*: 11 features, 90% recognition rate) and [11] (multilayer parceptron, training(50%) / test(50%) data split, *sonar*: 10 features, 81.42% recognition rate), as the classifiers, selection techniques and performance measurements are considerably different. Nevertheless, the recognition rates for the obtained dimensionality reductions in experiments 1 and 2 leading to conclude that the proposed method outperforms the Recursive Feature Elimination and achieved a similar result to the LOO recognition rate based method. Also, it provided a better feature ranking, being

appropriate for application in real-world domain.

IV. Conclusions

In this study, a new feature subset selection algorithm for classification task using SVM was developed. The proposed method implements the sequential backward selection strategy, and the margin of SVM is the base to the evaluation criterion of selected features. The Confident Margin measurement was introduced as a new selection criterion, which provides a better approach to evaluate the quality of the subset.

The effectiveness of the method was verified through several experiments. Three datasets were used in the experiments, including an artificial data and two from the real-world domain. As the result, in term of recognition rate and dimensionality reduction, in most of the cases the proposed method achieved better performance than the other algorithms. Further analysis to the confident margin curve of the proposed algorithm shows that it has a similar behavior to the recognition rate curve achieved by this ranking criterion. This fact provides the possibility to obtain the best subset by monitoring the peak of the confident margin curve without directly calculating the classifier recognition rate.

Acknowledgment

The research of M.K. is supported by the Ministry of Education, Culture, Sports, Science and Technology, Government of Japan, and also by the grant from the Hori Information Science Promotion Foundation, Japan. The research of A.S.N is partially supported by the Grant-in-Aid for Private University High-Tech Research Center from Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] A. Jain and D. Zongker, "Feature selection: Evaluation , application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[3] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, J. Ares, Manuel, and D. Haussler, "Support vector machine classification of microarray gene expression data," University of California, Santa Cruz, Tech. Rep. UCSC-CRL-99-09, June 1999.

[4] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, J. Ares, Manuel, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Science*, vol. 97, no. 1, pp. 262–267, January 2000.

[5] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection - theory and algorithms," in *Proceedings of the 21st International Conference on Machine Learning (ICML04)*. New York: ACM Press, 2004.

[6] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

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

[8] T. Marill and D. Green, "On the effectiveness of receptors in recognition systems," *IEEE Transactions on Information Theory*, vol. 9, pp. 11–17, 1963.

[9] C. Blake and C. 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.

[10] C. Emmanouilidis, A. Hunter, and J. MacIntyre, "A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator," in *Proceedings of the 2000 Congress on Evolutionary Computation (CEC00)*. California: IEEE Press, 6-9 2000, pp. 309–316.

[11] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667 – 1671, December 2002.