Southeast Asian Journal of Sciences: Vol 1, No. 2 (2012) pp. 137-147

FEATURE SELECTION WITH RANKING SUPPORT VECTOR MACHINE VISUALIZATION

Nguyen Thi Thanh Thuy

Department of Information Technology
Posts and Telecommunications Institute of Technology (PTIT)
Hanoi, Vietnam
e-mail: thuyntt@ptit.edu.vn

Abstract

In this paper, we first consider an application of the Nomogram visualization technique, which is a well-known one for describing numerical relationships in a graph, to ranking support vector machine. And then we utilize it to construct a feature selection method for ranking problems. In order to represent each feature on the log odds ratio in the nomogram, we use a probabilistic ranking support vector machine. Its purpose is to map the ranking support vector machine outputs into a probabilistic sigmoid function whose parameters are trained by using cross-validation. The effectiveness of our proposal helps the analysts study the effects of predictive features. Evaluation of the performance of ranking support vector machine visualization on the OHSUMED datasets shows that the proposed method is effective in feature selection.

1 Introduction

Feature selection recently has gained increasing attention in the data mining field with many applications such as text mining, bioinformatics, sensor networks, etc. Its purpose is to select a subset of relevant features and also removes irrelevant and redundant features from the data to build robust learning models. There are many potential benefits of feature selection: facilitating data

Key words: Nomogram, visualization, SVM, ranking SVM, probabilistic ranking SVM.

visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, and defying the curse of dimensionality to improve prediction performance [7]. There are three main feature selection methods in the literature: *filters*, *wrappers* and *embedded methods*. The *filters* select features by ranking them with correlation coefficients (based on a statistical score). The *wrappers* assess subsets of variables according to their usefulness to a given predictor. And the *embedded methods* perform feature selection as part of the learning procedure and usually specify to given learning machines. The wrapper and filter methods are usually more efficient in computation than the embedded methods, because their feature selection is independent of the classification method. However, the embedded methods produce more accurate results in general because they take advantage of properties of the classification method to maximize the accuracy of feature selection [7], [9].

SVM-RFE *(Support Vector Machine - Recursive Feature Selection)* is an embedded feature selection algorithm based on support vector machine, which was recently proposed to select a relevant set of features for a cancer classification problem [8]. *Nomogram-based RFE feature selection* is a method in which a feature is more important when the length of its line in the nomogram representation is longer. Consequently, features having small effect are removed by computing their length in the nomogram representation. Because features with small effect may be noisy or redundant features which reduce the accuracy of the classifier.

Ranking support vector machine (Ranking SVM) [1] is the most favorite ranking method that was applied to various different applications [2], [3], [4]. Besides its various advantages, ranking SVM still has difficulty in intuitively presenting the classifier which is also the disadvantage of original SVM. Inspired by the nomogram based visualization for SVMs of Jakulin [5], we also proposed a method which intuitively presents the ranking SVM. In order to present a ranking SVM on a nomogram, we must use the posterior probabilities of the output of ranking SVM proposed in [6].

Our contribution in this paper is two-fold: firstly, we propose a nomogram based visualization method for ranking SVMs; and secondly, based on the nomogram presentation, we improve a nomogram-based RFE feature selection method for ranking problems.

The remaining of this paper is organized as follows: First, we briefly summarize an approach for visualization of Support vector machines in Section 2. Following is the nomogram-based RFE algorithm for eliminating irrelevant and redundant features which having the shortest length in the nomogram. In section 3, we propose a new approach that uses nomogram to visualize Ranking support vector machine. Experimental results and conclusions are described in Section 4 and Section 5, respectively.

2 Nomogram Visualization in Classification Problem

2.1 Nomogram Visualization for SVM

In this section, we briefly discuss how to visualize a Support Vector Machines (SVM) model with a method proposed by Jakulin in [5]. This approach employs logistic regression to convert the distance from the separating hyperplane into a probability, and then represents the effect of each predictive feature on the log odds ratio scale as required for the nomogram. The main advantage of this approach is that it captures a complete classification model in a single, easy to interpret graph and for all common types of attributes and even for non-linear SVM kernels.

Suppose that we have a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_1^l$ in which $\mathbf{x}_i$[1] is a feature vector in n dimensional feature space $\Re^n$ and $y_i \in \{+1, -1\}$ is the class label of $\mathbf{x}_i$. The distance from a sample $(\mathbf{x}_i, y_i)$ to the separating hyperplane of the SVM can be replaced by the decision function in the SVM as follows [10], [11]:

$$f(\mathbf{x}) = \sum_{1}^{M} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{1}$$

where $M(< l)$ is the number of support vectors, $\alpha_i > 0$ are the Lagrange multipliers for support vectors, b is bias, and $K(\mathbf{x}_i, \mathbf{x})$ is called kernel function, that returns a similarity between $\mathbf{x}_i$ and $\mathbf{x}$. Depending on positive or negative sign of $f(\mathbf{x})$, SVM classifier predicts the label of an unknown instance of the testing dataset.

In the case of linearly decomposable kernel with respect to each feature, the distance becomes:

$$f(\mathbf{x}) = \sum_{k=1}^{n} [\mathbf{w}]_k + b \tag{2}$$

and the weight vector is defined as:

$$[\mathbf{w}]_k = \sum_{i=1}^{M} \alpha_i y_i K(\mathbf{x}_{i,k}, \mathbf{x}_k) \tag{3}$$

where n is the number of features, $\mathbf{x}_k$ is kth feature of sample $\mathbf{x}$, and $\mathbf{x}_{i,k}$ is kth feature of ith support vector [5], [9].

According to the method presented in [12], the posterior probability that the sample $\mathbf{x}$ belong to the positive class (in binary classification problem) is calculated as:

$$P(y = +1|\mathbf{x}) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \tag{4}$$

[1] We denote the bold variables as vectors or matrices

The two parameters A and B are fitted using maximum likelihood estimation from a training set and found by minimizing the negative log likelihood function of the training data. To avoid overfitting, a cross-validation method is used.

After finding two parameters A and B, these symbols A, B, $\mathbf{w}$ and b can be rewritten as the intercept β_0 and the effect function β. The intercept β_0 is a constant delineating the prior probability in the absence of any features, and the effect function β maps the value of a feature for the instance $\mathbf{x}$ into a point score, and finally using the inverse link function maps these functions into the outcome probability for an instance. The nomogram is based upon one effect function for each feature. Each line in the nomogram corresponds to a single feature, and a single effect function. The mapping is as follows:

$$\beta_0 = Ab + B \tag{5}$$

$$[\beta]_k = A[\mathbf{w}]_k \tag{6}$$

Then, the posterior probability (4) can be rewritten as:

$$P(y = +1|\mathbf{x}) = \frac{1}{1 + \exp(\beta_0 + \sum_{1}^{n} [\beta]_k)} \tag{7}$$

2.2 Nomogram-based Recursive Feature Elimination (RFE)

SVM-RFE is an embedded feature selection algorithm based on SVM for a classification problem [8]. At first, this algorithm starts with all features. At every iteration, feature weights are obtained by learning the training dataset with the existing features and then a feature with minimum weight is removed from the data. This procedure continues until all features are ranked according to the removed order. Similar to this method, Nomogram-based RFE algorithm [9] is used to remove features that have a low effect on the prediction output. Table 1 show detailed nomogram-based RFE algorithm.

Nomogram-based RFE algorithm is implemented via 3-fold cross-validations. Initially, the selected feature list is set to null, the training subset of features (or surviving features) is the full set of features. At each iteration, we run 3-fold cross validations to get the accuracy with the current subset of the surviving features. This accuracy is compared to the stored best accuracy (initially, best accuracy = 0). If the accuracy is greater, the selected feature list is set to the current subset of the surviving features and update the best accuracy to the current accuracy. At the end of each iteration, we will eliminate one feature from the current subset of the surviving features. The eliminated feature is the one having the shortest length in the nomogram. To compute the length of each feature in the nomogram, we train an SVM model with the restricted

Table 1: Nomogram-based Recursive Feature Elimination algorithm

Inputs	Training samples $\mathbf{X}_0 = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l]^\top$ and their class labels $\mathbf{y} = [y_1, y_2, \ldots, y_l]^\top$
1	Initialize subset of surviving features $\mathbf{s} = [1, 2, \ldots, n]$, selected feature list $\mathbf{r} = []$
2	Initialize $max\_accuracy = 0$
3	while $(\mathbf{s} \neq [])$
4	Restrict training samples to the subset of surviving features $\mathbf{X} = \mathbf{X_0}(:, \mathbf{s})$
5	Perform 3 fold cross validation with the restricted samples and their class labels, and get the accuracy of this cross validation: $test\_accuracy$
6	if $(max\_accuracy < test\_accuracy)$: $\{\mathbf{r} = \mathbf{s}; max\_accuracy = test\_accuracy\}$.
7	Train the SVM model with the restricted samples and their class labels.
8	Compute the nomogram representation of the trained SVM model.
9	Compute the length of each feature's range on nomogram representation.
10	Find the feature giving the shortest length, assuming it is s.
11	Eliminate the feature with the shortest length: $\mathbf{s} = [1, \ldots, s-1, s+1, \ldots, length(\mathbf{s})]^\top$.
12	end while
Outputs	Best feature list $\mathbf{r}$.

samples (the current subset of the surviving features) and compute the nomogram representation from the SVM model. The next iteration is implemented with the new subset of the surviving features. The loop ends when the subset of the surviving features is empty.

3 Proposed SVM Visualization for Ranking Problem

In this section, we propose a ranking SVM visualization based on nomogram for ranking problem. Assume that there is an input space $X \subset \Re^n$, where n is the dimension. And assume that we are given a ranking dataset (detailed in

[1], [4], [6]).

$$\mathcal{D}' = \{x_i^{(1)} - x_i^{(2)}, z_i\}_1^h, \quad x_i^{(1)}, x_i^{(2)} \in X \text{ for } i = 1, \ldots, h \tag{8}$$

And the score of the ranking SVM function is expressed as:

$$f(\mathbf{x}) = \sum_1^M \alpha_i z_i K(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}, \mathbf{x}) + b \tag{9}$$

where $M(< h)$ is the number of support vectors in the ranking problem. When the kernel is linearly decomposable with respect to each feature, the distance becomes:

$$f(\mathbf{x}) = \sum_{k=1}^n [\mathbf{w}]_k + b \tag{10}$$

and the weight vector is defined as:

$$[\mathbf{w}]_k = \sum_{i=1}^M \alpha_i z_i K(\mathbf{x}_{i,k}^{(1)} - \mathbf{x}_{i,k}^{(2)}, \mathbf{x}_k) \tag{11}$$

where n is the number of features, $\mathbf{x}_k$ is kth feature of sample $\mathbf{x}$, and $(\mathbf{x}_{i,k}^{(1)} - \mathbf{x}_{i,k}^{(2)})$ is kth feature of ith support vector.

The posterior probability that the sample $(\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$ belong to the positive class, it means that $z = +1$ or $(\mathbf{x}^{(1)} > \mathbf{x}^{(2)})$, is calculated as:

$$P(\mathbf{x}^{(1)} > \mathbf{x}^{(2)} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{1}{1 + \exp\{Af(\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) + B\}} \tag{12}$$

The above posterior probability for an output of ranking SVM was proposed in [6] which also discussed how to find the two parameters A and B.

Similarity with the nomogram visualization with SVM, we convert A, B, $\mathbf{w}$ and b to the intercept β_0 and the effect vector β, and use these parameters to represent the line of the Log OR for the feature in a nomogram.

$$\beta_0 = Ab + B \tag{13}$$

$$[\beta]_k = A[\mathbf{w}]_k \tag{14}$$

Thus, the posterior probability (12) can be rewritten as:

$$P(\mathbf{x}^{(1)} > \mathbf{x}^{(2)} | \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{1}{1 + \exp(\beta_0 + \sum_{k=1}^n [\beta]_k)} \tag{15}$$

Here, we also use the nomogram based-RFE algorithm as the same in section 2 to eliminate irrelevant and redundant features which having the shortest

length in the nomogram. Instead of considering the input with the beginning classification training set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_1^l$, we run the algorithm with the ranking training set $\mathcal{D}' = \{x_i^{(1)} - x_i^{(2)}, z_i\}_1^h$. It means that, the training samples consist of all pairs $(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)})$ with their class labels z_i. Other steps in the algorithm are invariant. The output is the best feature list.

4 Experimental Results

We evaluate the performance of ranking support vector machine visualization on the OHSUMED datasets using LIBSVM [13] and VRIFA [14]. We test our nomogram based method with two kernel: linear and the localized radial basic function (LRBF) kernel, that is a nonlinear kernel which was proposed by B.H.Cho in [9]. Both of them are proved to be linearly decomposable kernels.

A linear kernel:

$$K(\mathbf{x}_i, \mathbf{x}) = {\mathbf{x}_i}^T \mathbf{x} \tag{16}$$

A LRBF kernel is the summation of each feature similarity:

$$K(\mathbf{x}_i, \mathbf{x}) = \sum_{k=1}^{N} \exp(-\gamma(\mathbf{x}_{i,k} - \mathbf{x}_k)^2) \tag{17}$$

OHSUMED dataset is available in the LETOR package [15]. OHSUMED dataset consists of 348,566 references and 106 queries, which is a subset of MEDLINE, a database on medical publications. It extracted 25 features (10 from title, 10 from abstract, and 5 from title + abstract). There are totally 16,140 query-document pairs with relevance judgments. The relevance degrees of documents with respect to each query are judged on three levels: definitely relevant, possibly relevant, or irrelevant.

Figure 1 and Figure 2 show the nomogram of a linear ranking SVM. In the Figure 1 the left panel shows the effect ranges of all features, with an input instance indicated by a dot and the right panel shows the probability map and the final probability output with that respective instance. In this figure we observe that the feature 20 has the widest range (that means the most important), whereas the features 5, 6, 7, 15, 16, 17, and 21 have effect ranges equal to zero that means they contribute none to the accuracy of the classifier. This is due to an observation that the values of those features in the dataset are all equal to each other (because we only read a certain query for ranking), so it makes the ranking dataset with all data pairs $(x_i^{(1)} - x_i^{(2)})$ at feature ith is equal to zero. Thus $z_i = 0$, then $[\mathbf{w}]_k$ in the formula (3) equal to zero, so the effect function in the formula (6) is equal to zero. These features are called noisy features. In the Figure 2, the left panel shows the effect ranges of the best subset of selected features, with an input instance indicated by a

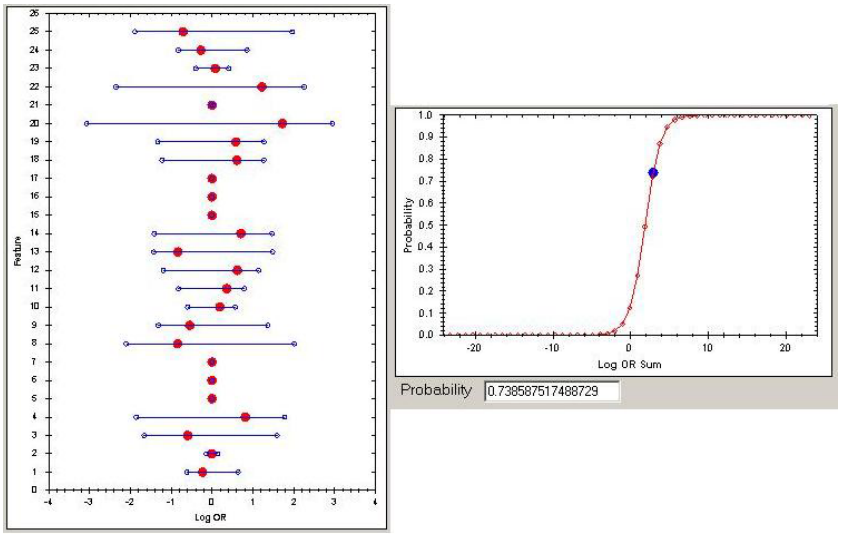

Figure 1: Nomogram visualization with a linear ranking SVM without using feature selection.

dot; the top-right panel shows the probability map and the final probability output with that respective instance; and the bottom-right panel shows the accuracy depending on the number of selected features. In this figure, the result shows only a subset of the selected features on the nomogram which makes the largest accuracy of the cross-validation. In addition, the figure also shows the accuracies of various feature selection. We observe that the best subset of selected features has 13 features (with highest accuracy is 0.7543) The eliminated features are: 1, 2, 5, 6, 7, 10, 11, 15, 16, 17, 21, and 23.

Figures 3 and 4 draw the respective nomogram of a LRBF kernel ranking SVM. Figure 3 draws the nomogram without feature selection. Figure 4 draws the best subset of selected features on the nomogram, and the accuracies depending on the various subsets of selected features. We observe that the best subset of selected features has 17 features that gives the highest accuracy = 0.7630. The eliminated features are: 2, 5, 6, 7, 15, 16 , 17, and 24.

5 Conclusion

In this paper, we proposed a nomogram based method to effectively visualize ranking support vector machines. Nomogram showed its effectiveness in presenting ranking SVM with many dimensions. More specifically, individual features are drawn vertically in a nomogram. Each line on the nomogram shows

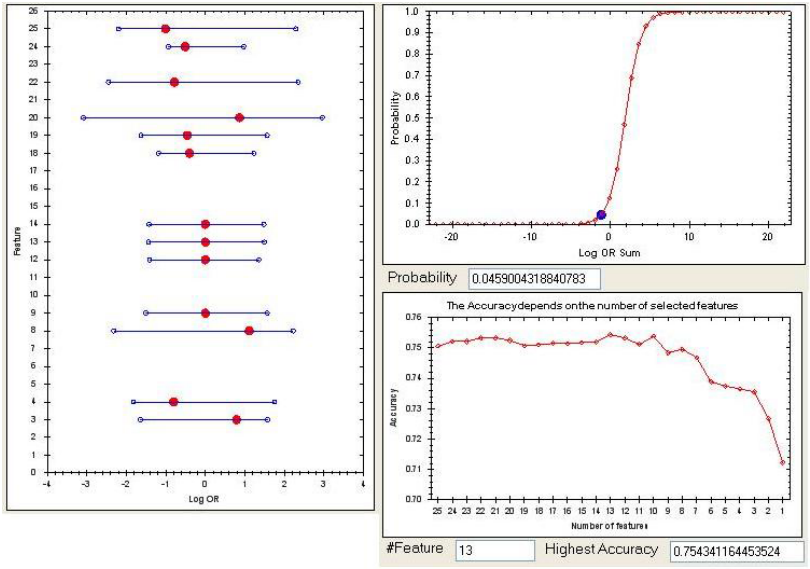

Figure 2: Nomogram visualization with a linear ranking SVM using feature selection.

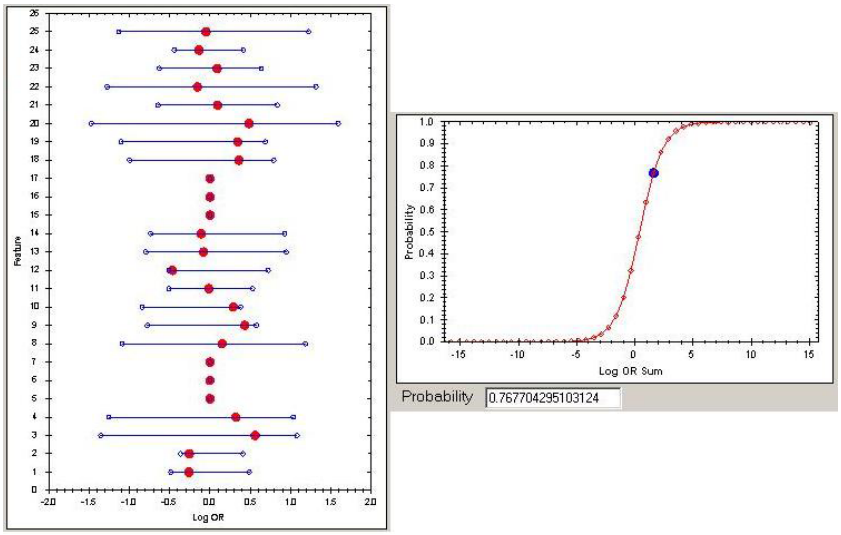

Figure 3: Nomogram visualization with a LRBF ranking SVM without using feature selection.

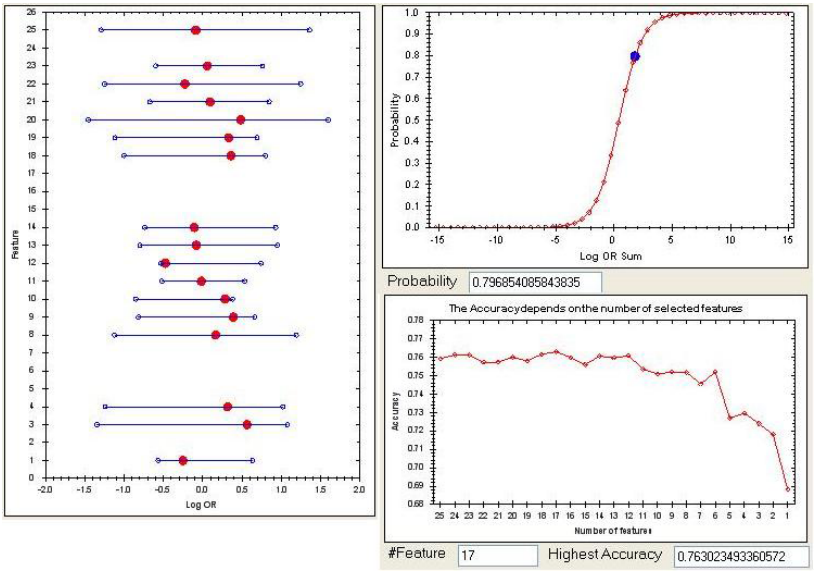

Figure 4: Nomogram visualization with a LBRF ranking SVM using feature selection.

the effect of one feature. In order to draw this nomogram, calibrated ranking SVM outputs [6] were used to calculate the effect function of features, and the ranking SVM was rewritten in the form of a generalized additive model. Through nomogram presentation, analysts can have an insight and study the effects of predictive factors. Furthermore, using nomogram presentation, we improved a nomogram based-RFE algorithm for a ranking SVM. The proposed feature selection technique showed its robustness in eliminating noisy and redundant features, and improved the overall accuracy. In the experiment, we drew nomograms with both linear and nonlinear (LRBF [9]) kernels which are both linearly decomposable.

Acknowledgment

We would like to thank anonymous reviewers, Prof. Dr. Tran Dinh Que and Dr. Ngo Anh Vien for their helpful suggestions and comments on our work.

References

[1] R. Herbrich, T. Graepel and K. Obermayer, *Large Margin Rank Boundaries for Ordinal Regression*, Advances in Large Margin Classifiers, MIT Press,

pp.115-132, 2000.

[2] H. Yu, *SVM Selective Sampling for Ranking with Application to Data Retrieval*, The eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp.354-363, ACM New York, 2005.

[3] H. Yu, Y. Kim and S. W. Hwang, *RVM: An Efficient Method for Learning Ranking SVM*, Technical Report, Department of Computer Science and Engineering, Pohang University of Science and Technology (POSTECH), Korea, http://iis.hwanjoyu.org/rvm, 2008.

[4] Y. Cao, J. Xu, T. Y. Liu, H. Li, Y. Huang and H.-W. Hon, *Adapting ranking SVM to Document Retrieval*, ACM SIGIR'06, pp.186-193, 2006.

[5] A. Jakulin, M. Mozina, J. Demsar, I. Bratko and B. Zupan, *Nomograms for visualizing support vector machines*, The eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp.108-117, 2005.

[6] N. T. T. Thuy, N. A. Vien, N. H. Viet and T. C. Chung, *Probabilistic Ranking Support Vector Machine*, ISNN (2), pp. 345-353, 2009.

[7] I. Guyon and A. Elisseeff, *An Introduction to Variable and Feature Selection*, Machine Learning Research, vol.3, pp.1157-1182, 2003.

[8] I. Guyon, J.Weston, S. Barnhill, M. D., and V. Vapnik, *Gene selection for cancer classification using support vector machines*, Machine Learning, vol.46, pp.389-442, 2002.

[9] B. Cho, H. Yu, J. Lee, Y. Chee, I. Kim and S. Kim, *Nonlinear Support Vector Machine Visualization for Risk Factor Analysis using Nomograms and Localized Radial Basis Function Kernels*, IEEE Transactions on Information Technology in Biomedicine, vol.12, pp.247-256, 2008.

[10] V. N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, 1998.

[11] C. J. C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery, Vol.2, pp.121-167, 1998.

[12] J. C. Platt, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*, Advances in Large Margin Classifiers, pp.61-74, MIT Press, 1999.

[13] C. C. Chang and C. J. Lin, *LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval*, LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[14] N. A. Vien, N. H. Viet, T. C. Chung, H. Yu, S. Kim, B. H. Cho, *Vrifa: a nonlinear svm visualization tool using nomogram and localized radial basis function (lrbf) kernels*, In: CIKM, pp.20812082, 2009.

[15] T. Y. Liu, J. Xu, T. Qin, W. Xiong and H. Li, *LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval*, The Learning to Rank workshop in the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR, 2007.