International Journal of Fuzzy Logic and Intelligent Systems, vol. 7, no. 2, June 2007 pp. 91-95

A Web Recommendation System using Grid based Support Vector Machines

Sung-Hae Jun

Department of Bioinformatics & Statistics, Cheongju University, 360-764 Chungbuk, Korea

Abstract

Main goal of web recommendation system is to study how user behavior on a website can be predicted by analyzing web log data which contain the visited web pages. Many researches of the web recommendation system have been studied. To construct web recommendation system, web mining is needed. Especially, web usage analysis of web mining is a tool for recommendation model. In this paper, we propose web recommendation system using grid based support vector machines for improvement of web recommendation system. To verify the performance of our system, we make experiments using the data set from our web server.

Key words : Web Recommendation System, Support Vector Machines, Grid Search

1. Introduction

Web recommendation system has been the most discussed and least understood aspect of the Internet. Many issues about how the recommendation system changes traditional web business models have been discussed[3],[9]. But there is little solution about the recommendation system. According as Internet commerce gives rise to new kinds of business models, the recommendation system is a good example for internet shopping malls. In our works, we use the web usage mining to construct an effective recommendation system. Web mining can be broadly defined as the discovery and analysis of useful information from the world wide web[1],[2],[7]. Generally, web mining tasks can be classified into three categories which are web content mining, web structure mining, and web usage mining[5],[11]. Among them, the web usage mining is mostly to analyze web log data. The web log records contain much collection of hyperlink information and the usage transactions of web page access. The size of web log data is very large, but web log data are very sparse. So we have serious difficulties for web mining. It is very difficult to estimate the dependencies of all web pages from spare web log data. We have found that support vector machines(SVMs) are efficient approaches for analyzing the sparse data because of its ε-insensitive loss function[13]. Using SVMs as a missing value imputation, the spare data set is changed to complete data set[8]. Our previous research provided a useful strategy for analyzing sparse data like web log data. But, SVMs have some problems in data analysis. One of the problems is subjective determination of the parameters of SVMs. The parameters affect the result performance of SVMs. However, in many cases, the parameter determinations are depended on the arts of researchers. So, in this paper, we use grid search for objective and automatic determinations of SVMs parameters. Using grid based SVMs, we are able to construct an effective web recommendation system. By experimental results using web log data from our web server[14], we verify improved performances of our recommendation system.

Manuscript received Mar. 3, 2007; revised Apr. 13, 2007.

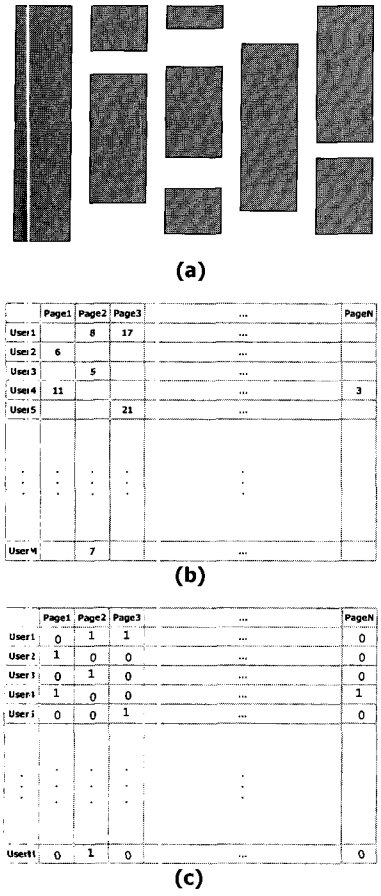
2. Web Recommendation System

Web recommendation systems offer website visited users proper web pages efficiently[9]. The recommendation systems are built using web usage mining by analyzing web log data. The web log data comes from log file of website. The sparseness of web log file has been a problem of web usage mining. This is occurred by several reasons. Frequently, it happens when the not visited web pages are much larger than the visited web pages in web sites. The click stream data of cleaned web log are very sparse. So, we have a difficulty of web log analysis as web usage mining with web information recommendation, next web page prediction, and web page duration time forecasting. The click stream data with sparseness is hard to analyze by general methods as regression, imputation methods, and others[5]. In this case, SVMs are very useful tools for analyzing sparse data[8]. But SVMs have had some problems which are optimal selection of the parameters of SVMs. Therefore, we have needed to solve the problems of SVMs. To settle the problems, we use grid based SVMs in this paper.

3. Grid based SVMs for Web Recommendation System

We propose two models for constructing an efficient web recommendation system. In approach of web usage mining,

firstly we predict the duration time of web page. Also, we estimate the visited probability of web page. Following figure shows our approach of recommendation system.



(a)

	Page1	Page2	Page3	...	PageN
User1		8	17	...	
User2	6			...	
User3		5		...	
User4	11			...	3
User5			21	...	
...	...	...	...	...	...
UserM		7		...	

(b)

	Page1	Page2	Page3	...	PageN
User1	0	1	1	...	0
User2	1	0	0	...	0
User3	0	1	0	...	0
User4	1	0	0	...	1
User5	0	0	1	...	0
...	...	...	...	...	...
UserM	0	1	0	...	0

(c)

Fig. 1. Our Web Usage Mining for Recommendation

(a) of above figure shows missing type of web log data. This missing type is very difficult to analyze. In this paper, using grid based SVMs, we are able to analyze our web log data efficiently. After preprocessing web log file, the click-stream data table are constructed as (b) of figure 1. the low and column represent the visited user and web pages respectively. Similar to (b), (c) is also click-stream data set. But, each cell of (c) has 0 or 1. the value 1 shows the web page is visited. The value of not visited web page is 0. In the click-stream data, we perform web usage mining by the following SVMs. Our given training data consist of N pairs, $(x_1, y_1), \ldots, (x_N, y_N)$, where x denotes the input patterns and y is target variable. In SVMs for regression with ε-insensitive loss function, our goal is to find a function f(x) that has at most ε-deviation from the actually obtained targets yi for all the training data, and at the same time, is as flat as possible[12]. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this. ε-insensitive loss function is defined as,

$$M(y, f(x,\alpha)) = L(|y - f(x,\alpha)|_\varepsilon) \tag{1}$$

We denote,

$$|y - f(x,\alpha)|_\varepsilon = \begin{cases} 0, & if\ |y - f(x,\alpha)| \le \varepsilon, \\ |y - f(x,\alpha)| - \varepsilon, & o.w. \end{cases} \tag{2}$$

α is a positive constant. The loss is equal to 0 if the discrepancy between the predicted and the observed values is less than ε. The case of linear function f is described.

$$f(x) = < w, x > + b \tag{3}$$

where $< \cdot, \cdot >$ denotes the dot product. For SVMs, the Euclidean norm $\| w \|^2$ is minimized. Formally this problem can be written as a convex optimization problem by requiring[13]. Analogously to the loss function in [13], we introduce slack variables ξ_i, ξ_i^* to copy with otherwise infeasible constraints of the optimization problem.

$$\text{minimize } \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{4}$$

$$\text{subject to } \begin{cases} y_i - < w, x_i > - b \le \varepsilon + \xi_i \\ < w, x_i > + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{5}$$

The constant C>0 determines the trade off between the flatness of f and the amount up to which deviations larger than ε are tolerated. Using a standard dualization method utilizing Lagrange multipliers, the parameters are determined from equation (4) and (5). In SVMs, we face on some problems which affect the results of SVMs. One of the problems is subjective determination of kernel parameter and regularization constant by the arts of researchers. But, the determinations are objective for improving SVMs. Next, to settle the problem of SVMs, which are the determinations of kernel parameters and regularization constant of SVMs, we use grid searching as the following figure.

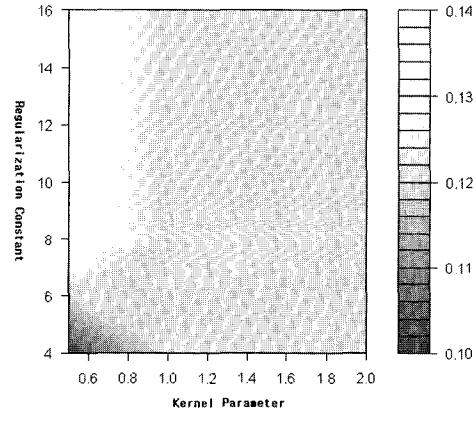


Fig. 2. Grid Search for the parameters of SVMs

A searching interval of kernel parameter of SVMs is shown the x axis in the above figure. Y axis of the figure is represented a searching interval of regularization constant. In the figure 3,

dark area has optimal value for the parameters of SVMs. The vertical bar shows misclassification rates according to the dark density. In this paper, for the sparseness elimination from click stream data, grid based SVMs as the missing value imputation approach is used. This has a good performance for sparse data analysis because of its ε-insensitive loss function[13]. This is able to transfer incomplete data of (b) figure 1 into complete for web usage mining. Each cell of table in (b) of figure 1 contains a duration time of user accessing. The sparseness of cells is very serious. Therefore, general preprocessing methods like the missing value imputation methods as multiple imputation method are not suitable to solve this problem. This fact will be verified next experiments. The duration time of ith web page is estimated as following equation.

$$\hat{T}_{Pagei} = f(T_{Page1},\ldots,T_{Page(i-1)},T_{Page(i+1)},\ldots,T_{PageN}) \quad (6)$$

In above equation, T_{PageK} was defined the duration time of page K by user accessing and $\hat{T}_{Pagei}$ was defined the estimated duration time. $\hat{T}_{Pagei}$ was computed by the ESVR method of (N-1) pages, $T_{Page1},\ldots,T_{Page(i-1)},T_{Page(i+1)},\ldots,T_{PageN}$ out of i page. Therefore we predict duration time of each web page using estimating missing cells.

Proposed web recommendation system is consisted of web usage mining and campaign feed-back. The process of the system is performed from user accesses to user feed-back. In the following figure, the detailed specifications of our system are illustrated.

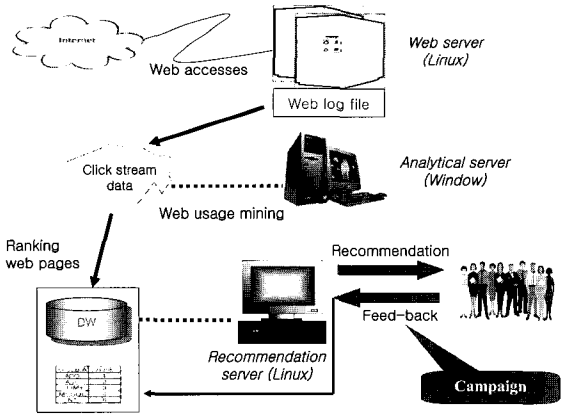


Fig. 3. Proposed Web Recommendation System

The recommendation system has three servers which are web server, analytical server, and campaign server. The web server offers web contents to accessed users. Grid based SVMs and campaign feed-back are performed in analytical and campaign servers respectively. According to the role partitions of servers, we can reduce the loads of server.

4. Experimental Results

To verify the performance of proposed web recommendation system, we use data set from our web server[14]. Also, R-project(e1071) package are used for web usage mining in proposed recommendation system[15]. In our experiments, the MSE(mean squared error) and the Lift value are used for the performance measure[4]. The MSE is defined as following.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(T_i - O_i)^2 \quad (7)$$

where, Ti is the ith target variable(known) and Oi is the ith predictive variable(unknown). N is the size of data set. The smaller the value of MSE is, the better the performance of the method is.

Next the Lift is a measurement of how much better the model predicted results for a given case set over what would be achieved through random selection. Lift is typically calculated by dividing the percentage of expected response predicted by the data mining model by the percentage of expected response predicted by a random selection. For example, suppose that 2% of the customers mailed a catalog without using the model would make a purchase. However, using the model to select catalog recipients, 10% would make a purchase. Then the lift is 10/2 or 5. Lift may also be used as a measure to compare different intrusion detection models. Since lift is computed using a data table with actual outcomes, lift compares how well a model performs with respect to this data on predicted outcomes. Lift indicates how well the model improved the predictions over a random selection given actual results. Lift allows a researcher to infer how a model will perform on new data. Generally the Lift value is defined as the following[4].

$$LV = \frac{\%response}{LV_{BL}} \quad (8)$$

In the above equation, the %response was percentage of the number of correctly predicted attacks using constructed model and the LVBL was the base line lift value which is the predicted result by random selection without modeling. The Lift value is 2, this means that using a model we are 2 times more likely to get a success than if we chose randomly without a model.

In this section, we show the experimental result using our web server data. This is the web log file of our laboratory web site[14]. Summary information of our experimental data is shown in the following table.

Table 1. Summary information of web log data

Attributes	Value range
IP address	2,000 (users)
Web page	45 (pages)
Duration time	0~1000 (seconds)

In above table, the numbers of users which are represented by IP address and web pages are 2000 and 45 respectively. The one-third of given data for the validation and the other two-thirds for training are used 10]. That is, in the above, the IP address is the index of user accessing to web site. The web page represents each web page containing the descriptive contents of each item in the web site. A user accesses a web page with duration times between 0 to .000 seconds.

In this experiment, we verify grid based SVMs by two supervised learning approaches which are regression and classification. In regression experiment, nonlinear regression, and traditional SVR methods with polynomial, RBF(radial basis function), and two layers MLP(multi-layers perceptron) kernels are compared with grid based SVMs for regression[6]. The experimental result is shown in the following table.

Table 2. Performance Comparison in Regression

Method		MSE	LV
Nor linear regression		3.58	2.1
SVR	Polynomial	2.42	2.7
	RBF	2.20	3.3
	Two-layer MLP	2.78	2.6
Grid based SVM for regression		1.96	3.8

In this result, the MSE values of testing data are com-putted. The MSE of grid based SVMs is the smallest in the comparative methods.

Next, we compare grid based SVMs with logistic regression, and traditional SVM methods with polynomial, RBF, and two layers MLP kernels in the classification case. Following table shows the experimental result.

Table 3. Performance Comparison in Classification

Method		Accuracy
Logistic regression		0.80
SVM	Polynomial	0.93
	RBF	0.89
	Two-layer MLP	0.91
Grid based SVM for classification		0.95

Similar to the regression result, this result also shows the accuracy of grid based SVMs for classification is the best in the comparative methods. Therefore, we find improved performance of our web recommendation system.

5. Conclusions

In this paper, we propose a web recommendation system. The system is constructed by web usage mining which uses grid based SVMs. That is, we use grid based SVMs as methods for web usage mining which contain regression and classification. In our experiment, we verify improved performance of our recommendation system by MSE, Lift value, and accuracy. Compared with popular methods, we find usability of proposed recommendation system.

References

[1] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases", *Proceeding of the ACM SIGMOD International Conference on Management of Data*, 1993.

[2] R. Cooley, P. N. Tan, J. Srivastava, "Discovery of interesting usage patterns from web data", *Technical Report TR 99-022, University of Minnesota*, 1999.

[3] D. Fisher, K. Hildrum, J. Hong, M. Newman, M. Thomas, R. Vuduc, "SWAMI: A Framework for Collaborative Filtering Algorithm Development and Evaluation", *Proceeding of SIGIR 2000*, ACM Press, 2000.

[4] P. Giudici, *Applied Data Mining*, Wiley, 2003.

[5] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.

[6] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.

[7] J. S. Jang, S. H. Jun, K. W. Oh, "Fuzzy Web Usage Mining for User Modeling", *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 2, no. 3, pp. 204-209, 2002.

[8] S. H. Jun, "Web Usage Mining Using Support Vector Machine", *Lecture Note in Computer Science*, vol. 3512, pp. 349-356, 2005.

[9] P. Kazienko, P. Kuzminska, "The influence of indirect association rules on recommendation ranking lists Intelligent Systems Design and Applications", *Proceedings of 5th International Conference on ISDA*, pp. 482 – 487, 2005.

[10] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[11] P. Resnick, N. Lacovou, M. Suchak, P. Berfstrom, J. Riedl, "GroupLens: An Open Architecture for collaborative filtering of Netnews", *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, 1994.

[12] A. J. Smola, *Regression estimation with support vector learning machines*, Master's thesis, Technische University, 1996.

[13] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons,

Inc. 1998.
[14] http://delab.cju.ac.kr
[15] http://www.r-project.org.

Sung-Hae Jun
He received the BS, MS, and PhD degrees in department of Statistics, Inha University, Korea, in 1993, 1996, and 2001. Also, He received PhD degree in department of Computer Science, Sogang University in 2007. He is currently Assistant Professor in department of Bioinformatics & Statistics, Cheongju University, Korea. He has researched statistical learning theory and evolutionary algorithms.

Phone : +82-43-229-8205
Fax : +82-43-229-8432
E-mail : shjun@cju.ac.kr