Proceedings of KFIS Autumn Conference 2006 Vol. 16, No. 2.

Hybrid Internet Business Model using Evolutionary Support Vector Regression and Web Response Survey

Sung-Hae Jun
Dept. of Bioinformatics & Statistics, Cheongju University
360-764 Chungbuk, Korea
shjun@cju.ac.kr

Abstract

Currently, the nano economy threatens the mass economy. This is based on the internet business models. In the nano business models based on internet, the diversely personalized services are needed. Many researches of the personalization on the web have been studied. The web usage mining using click stream data is a tool for personalization model. In this paper, we propose an internet business model using evolutionary support vector machine and web response survey as a web usage mining. After analyzing click stream data for web usage mining, a personalized service model is constructed in our work. Also, using an approach of web response survey, we improve the performance of the customers'satisfaction. From the experimental results, we verify the performance of proposed model using two data sets from KDD Cup 2000 and our web server.

I. Introduction

The internet business models(iBMs) have been the most discussed and least understood aspect of the web. Many issues about how the web changes traditional business models have been discussed[4]. But there is little solution of exactly what this means. Internet commerce will give 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 a good iBM. Web mining can be broadly defined as the discovery and analysis of useful information from the world wide web[2],[3],[5],[7],[8]. The size of web log data is very large, but web log data are very sparse. So, in this paper, combining the evolutionary computing into SVR, we propose an ESVR(evolutionary SVR) for web usage mining. For an efficient iBM, we consider how visitor behavior on a website can be predicted by analyzing existing data on the order in which the site's web pages are visited. In addition to, we think over the web response survey for improving the performance of our iBM. This iBM is called Hybrid iBM(HiBM) in this paper. Our model offers a good result in spare web log data. In experiments using KDD Cup 2000 data and the web log data of our web server, we verified the performance of our work[10],[11].

II. Hybrid Internet Business Model

2.1 An Evolutionary Support Vector Regression for Web Usage Mining

Evolutionary computing is a special type of computing, which draws inspiration from the process of natural evolution. The fundamental of

evolutionary computing relates powerful natural evolution to a particular style of problem solving, that of trial and error[1]. Environment, individual, and fitness of the basic evolutionary computing were linked respectively problem, candidate solution, and quality of the natural evolution to problem solving. In this section, we proposed our ESVR(evolutionary SVM for regressive model). Genetic algorithm(GA) has provided a analytical method motivated by an analogy to biological evolution[1]. General GA computes the fitness of given environment where is fixed. Distinguished from traditional GA, co-evolving approach is evolutionary mechanism of the natural world with competition or cooperation. The organism and the environment including organism evolve together. We applied not cooperation but competition to our proposed co-evolutionary model. Our competitive co-evolving approach used host-parasites co-evolution. The host and parasites were used for modeling ESVR and training data set. Our ESVR and training data set were considered as the organism and the environment including it. That is, the evolving ESVR was followed the evolution of host. The initial parameters for ESVR model were determined as uniform random numbers from -1 to 1. The fitness function of ESVR was the inverse form of the squared error between real and predict values as following.

$$f_{host}(x) = \frac{C}{\sum_{i=1}^{F}\sum_{j=1}^{N_{out}}(o_{ij}(x)-t_{ij})^2} \quad (1)$$

In above equation, t was the value of known target variable and $o$was computed output value for prediction. C was a constant. The F and N_{out} were the numbers of patterns and items. Next, the training of given data set was performed by evolving parasites. The evolution of training data was performed to retain larger training errors. So the fitness function for training data set was inverse of the fitness function of ESVR model as the following.

$$f_{parasites}(x) = \sum_{i=1}^{D}\sum_{j=1}^{N_{out}}(o_{ij}(x)-t_{ij})^2 \quad (2)$$

The D and N_{out}were the numbers of patterns and items in the above equation. Our evolutionary approaches of ESVR and training data set were competitive. In other words, proposed model was two different groups'competitive co-evolving. One was the parasites' evolution of given training data set. Another was the host's evolution of ESVR. The ESVR model and training data set were respectively evolved. During evolution for weight optimization of ESVR, the competitive co-evolving was occurred between evolving SVM model and evolving training data set. In this place, our model used co-evolutionary computation instead of Lagrange multipliers of traditional SVM for parameter optimization.

2.2 Ranking Web Pages

In this paper, for the sparseness elimination from click stream data, the missing value imputation approach was used. Our imputation method is ESVR. This has a good performance for sparse data analysis because of its ε-insensitive loss function[6]. And this satisfies conditions for consistency of risk minimization principle[9]. What is more, ESVR had an evolutionary approach to solve local optima of general SVR. This must be complete for web usage mining.

2.3 Web Response Survey and HiBM Building

Combined with web response survey, our ESVR model is to be HiBM. The following figure shows the hierarchy of HiBM.

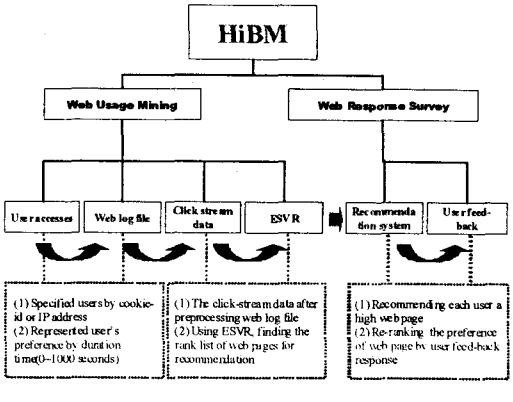

Fig. 1. HiBM hierarchy

HiBM is consisted of web usage mining and web response survey modules. The process of HiBM is performed from user accesses to user feed-back. In the above figure, the detailed specifications of HiBM are illustrated. We can sketch the system of HiBM in the following figure. The HiBM system has three servers which are web server, analytical server, and survey server. The web

server offers web contents to accessed users. The ESVR and e-mail response are performed in analytical and survey servers respectively. According to the partition of server's role, we can reduce the loads of server.

III. Experimental Results

Firstly we show the result of experiment using KDD Cup 2000 data. This is the web log file of real internet shopping mall(gazelle.com). The data size is 1.2 GB. The numbers of users(represented cookie-id) and web pages(represented assortment-id) are 13109 and 269 respectively. The one-third of given data for the validation and the other two-thirds for training are used[9]. That is, in the above, the cookie-id is the index of user accessing to web site. The assortment-id represents each web page containing the descriptive contents of each item in the shopping mall. A user accesses a web page with duration times between 0 to 1000 seconds.

In this experiment, the P-score(propensity score), MCMC(Markov Chain Monte Carlo), nonlinear regression, and traditional SVR methods with polynomial, RBF(radial basis function), and two layers MLP(multi-layers perceptron) kernels are compared with our ESVR[5],[6]. The comparative methods with our method have been published by their good performances in the web usage mining. The experimental result is shown in the following table.

Table 1. Result of evaluation: KDD Cup

Methods		MSE (total)	MSE (10%)
Multiple Imputation	P-score	3.10	2.38
	MCMC	2.36	1.98
Nonlinear regression		2.58	2.17
SVR	Polynomial	2.11	1.45
	RBF	1.69	1.21
	Two-layer MLP	2.03	1.38
ESVR		1.42	0.98

In this result, the MSE values of total and upper 50% of testing data are com-putted. The MSE of SVR is smaller than multiple imputation and nonlinear regression methods. But, the MSE of ESVR is the smallest in the comparative models. Therefore, we find the ESVR has a good performance. Next we use the web log data of our web server for another experiment. This size is smaller than KDD Cup data in previous experiment. Our web server has been used for informing our laboratory. It has contained members' profiles, lecture notes, research information, and so forth. Each user is identified by session id in the above table. The content id represents each web page for research and lecture contents. In this experiment, if a user accesses same web page twice and more, we sum the duration times about the web page. The duration time of not accessed web page is set 0. Same as previous experiment, we use the 66.7% of given data for training and the other 33.3% for validation[5]. Also, in the same way, we compare our method with the competitive methods. The following table shows the experimental result.

Table 2. Result of evaluation: Our web server

Methods		MSE (total)	MSE (10%)	Lift value
Multiple Imputation	P-score	4.09	2.85	2.0
	MCMC	2.87	2.19	2.6
Nonlinear regression		3.16	2.43	2.3
SVR	Polynomial	2.15	1.99	2.5
	RBF	1.98	1.10	3.0
	Two-layer MLP	2.50	2.01	2.4
ESVR		1.32	0.83	4.5

We fid the MSE of ESVR is the smallest in the comparative methods. In addition, we perform the web response survey by e-mail of the accessed users to our web server. They respond to satisfaction about the recommended web pages from the server. We use Lift value to verify the performance of web response. The Lift value of ESVR is the highest among the competitive methods. So, we know the improved performance of proposed method.

IV. Conclusion

In this paper, we propose an internet business model called HiBM. The model is consisted of the web usage mining and web response survey. Also, we construct an ESVR as new method for web usage mining. Further, using the web response

sur-vey, we improve the performance of the proposed method. In our work, we verify the performances of ESVR by the MSE and e-mail response by the Lift value. Compared with popular methods, we know the performance of HiBM is efficient. Our future work will be to develop the ESVC(evolutionary SVM for clustering) for our HiBM.

IV. 참고문헌

1. Casella, G., Berger, R. L.: Statistical Inference, Duxbury Press. (1990)
2. Dillman, D. A.: Mail and Internet Surveys - The Tailored Design Method, Wiley, (2000)
3. Eiben, A. E., Smith, J. E.: Introduction to Evolutionary Computing, Springer, (2003)
4. Giudici, P.: Applied Data Mining, Wiley, (2003)
5. Han, J., Kamber, M.: Data Mining Concepts and Techniques, Morgan Kaufmann, (2001)
6. Haykin, S.: Neural Networks, Prentice Hall, (1999)
7. Jang, J. S., Jun, S. H., Oh, K. W.: Fuzzy Web Usage Mining for User Modeling, International Journal of Fuzzy Logic and Intelligent Systems, vol. 2, no. 3, pp. 204-209, (2002)
8. Jun, S. H.: Web Usage Mining Using Support Vector Machine, Lecture Note in Computer Science, vol. 3512, pp. 349-356, (2005)
9. Vapnik, V. Z.: Statistical Learning Theory, John Wiley & Sons, Inc. (1998)
10. http://delab.cju.ac.kr
11. http://www.ecn.purdue.edu/KDDCUP