Online Outlier Detection System for Learning Time Data in E-Learning and Its Evaluation.

Conference Paper · January 2004

Source: DBLP

CITATIONS
18

READS
85

1 author:

Maomi Ueno
The University of Electro-Communications
121 PUBLICATIONS **372** CITATIONS

SEE PROFILE

ONLINE OUTLIER DETECTION SYSTEM FOR LEARNING TIME DATA IN E-LEARNING AND IT'S EVALUATION

Maomi Ueno
Nagaoka University of Technology
1603-1 Kamitomioka Nagaoka, Niigata, 940-2188,
Japan
ueno@kjs.nagaokaut.ac.jp

Abstract

Recently, distance education by using e-Learning has become popular in actual educational situations. However, there is a problem that the instruction strategy tends to be one way, and so it sometimes makes the learners bored comparing with usual instruction methods. This paper proposes a method of on-line outlier detection of learners' irregular learning processes by using the learners' response time data for the e-Learning contents. The unique features of this method are as follows: 1.It proposes an outlier detection method by using Bayesian predictive distribution. 2. It is available for small sample, 3.It provides an unified statistical test method of the various statistical test by changing the hyper-parameters, and it provides accurate test results than one of the traditional methods. 4. On-line outlier detection is realized on WWW. 5.It assists two ways instruction by using data mining results for the learners' learning processes. 6. The outlier statistics is estimated by considering both the students' abilities and contents' difficulties. This paper evaluated the efficiency of the proposal , and the results show the efficiency of the system.

Key Words

e-learning, LMS, Data Miming, Learning Histories Data Base

1. Introduction

Recently, distance education using e-Learning has become popular in actual educational situations. However, there is a problem that the instruction strategy tends to be one way, and so it sometimes makes the learners bored comparing with usual instruction methods. To solve these problems, it is an important task to develop a new methodology of instruction based on the e-Leaning.

On the other hand, it is easy to get huge learning histories data which has been saved as log data in the e-Learning . In this case, it is important how we save this data or how we utilize this data. Some researches has been proposed from this view point. Kawamura, S., (1998)[1] and Matsumoto,J., (1999)[2] proposed a clustering method for learners using e-Learning log data. Matsui, T and Okamoto, T.(2000) [3] proposed a data-mining method which constructs a tree using the ID3 method. This paper proposes a method of online outlier detection of learners' irregular learning processes using the learners' response time data for the e-Learning contents. Many outlier detection techniques have been proposed, for example,. See [4],[5],[6],[7], and [8]. However, there are the following problems in the case of applying the traditional outlier detection techniques to the detection problem of irregular learning processes in e-Learning situations.

- If a learner provides irregular learning processes from beginning of learning, then the regular learning processes will be regarded as a irregular processes.
- The traditional techniques assume that all data in the time series depends on the same task. However, in educational situations, the tasks (= contents) in the time series are respectively different in the sense of difficulty.
- In the traditional techniques, the criteria which specify the outlier are not clear in the sense of statistics.

In consideration of these problems, this paper proposes a new outlier detection method to detect a learner's irregular learning processes in e-learning. The unique advantages of this technique are as follows:

- The proposed method can combine the prior knowledge about the contents properties using Bayesian approach, it can avoid to regard regular processes as irregular processes at the beginning of the learning processes.
- The proposed method employs the model which depends on task difficulties and learners abilities. The outlier statistics is estimated considering both the students' abilities and contents' difficulties, then it is efficient in the case of detecting irregular learning processes in e-leaning.
- The proposed methods derives an unified statistical test method of the various statistical method, and this has a clear mean and criterion in the sense of the statistical predictive distribution.

In addition, we developed a LMS (Learning Management System) with the on-line outlier detection system. This system supports an on-line outlier detection of learners' irregular learning processes and it also assists two ways instruction using data mining results for the learners' learning processes. The system was utilized

for actual classes. The results show the efficiency of the system.

2. LMS(Learning Management System)

The author has developed a LMS (Learning Management System) ([9] and [10]). In this session, the outline of the system will be introduced. The LMS consists of 1. Contents Presentation System (CPS), 2. Contents Database (CD), 3.Learning Histories Database (LHD), and 4. Data Mining System (DMS). The CPS integrates various kinds of contents and present the integrated information on the web page shown in Figure 1. Moreover, the system presents some test items which confirm learners' comprehension degree as soon as the contents has been completed . An example of a test presentation is shown in Figure 2.

The CD is a database which consists of various kinds of medias, text, jpeg, mpeg, and so on. The proposed platform monitors learners' learning processes and saves them as a log data in the LHD. First, teacher makes the contents concerned with his lecture, and saves them in the CD. Then, the CPS automatically integrates the contents, and presents them to learners. The learners can learn them through the internet. The learners' learning histories log data is saved in the LHD, and it is analyzed in the DMS. The DMS presents the feedbacks for the learners and the teacher respectively. The teacher can know information about learners' learning processes, and he can give some comments or instructions to the learners using e-mail.

3. Learning Log Database

The LMS monitors learners' learning processes and saves them as a log data in the LHD. The saved data

Figure 1. An example of e-Leaning instruction

Figure 2. An example of Test frame

consists of A) Contents ID, B) Learner ID, C) The number which the learner has learned the content, D) Test Item ID, E) Operation order ID, F) Operation ID which indicates what operation was done in the content, G) Date and Time ID which indicates the time and date of starting the operation, and H) Time ID which indicates time that it takes in the operation. The problem is how we efficiently use the huge amount of log data. The next section will propose one of utilization of this data, a method of online outlier detection of learners' irregular learning processes.

4. On-Line Statistical Outlier Detection

4.1. Data

The data which is used for the outlier detection is response time data as shown in Figure 3. The horizontal axis indicates the number of contents which a learner has learned, and the vertical axis indicates the response time for the content. From this data, we will discover irregular learning processes. For this purpose, this paper proposes a new method to detect learner's irregular learning processes. The main idea is to derive a Bayesian predictive distribution of a new data x_{n+1} given the learner's learning processes data $x_1, \cdots, x_n$, and provides a test for outlier detection of the new data.

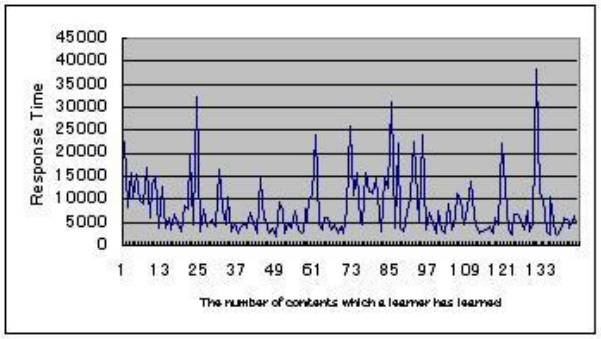


Figure 3. Response time data for the contents

4.2. Model

In this section, a Bayesian predictive distribution of a new data x_{n+1} given the learner's learning processes data $x_1, \cdots, x_n$ will be derived. Let t_{ij} be a learner *j*'s response time for the *i*-th content, and let consider the following linear equation

$$x_{ij} = \frac{t_{ij} - \bar{t}_i}{s_i} = \mu_j + e_j \quad , \quad (1)$$

A Bayesian predictive distribution of a new data x_{n+1} given the learner's learning processes data $x_1, \cdots, x_n$ can be derived as follows;

$$p(x_{n+1} \mid X) = \int\int p(x_{n+1} \mid \mu, \sigma^2) p(\mu, \sigma^2 \mid x_1, \cdots, x_n) d\mu d\sigma^2 \qquad (2)$$

$$= \left(1 + \left[\frac{(x_{n+1} - \mu_*)}{\sqrt{\frac{n_0 + n + 1}{(n_0 + n)\nu} \lambda_*^{\,2}}} \right]^2 / \nu \right)^{-\frac{\nu+1}{2}}$$

where

$$t = \frac{(x_{n+1} - \mu_*)}{\sqrt{\frac{(n_0 + n + 1)}{(n_0 + n)\nu} \lambda_*^{\,2}}} \qquad (3)$$

then t follows t distribution with degree of freedom $\nu = n_0 + n - 1$.

Here, μ_* indicates the hyper parameter of the prior distribution, which is the mean parameter of the normal distribution. The value of μ_* is determined the mean of data x. In this case, the data x is standardized with mean zero and standard deviation one, then μ_* is zero.This prior can combine the prior knowledge about the content with the outlier detection. For examples, even if a learner provides some irregular learning processes at the beginning of the learning, this method does not regard the processes as regular processes.

4.3. Outlier Detection

From (3), we can detect outlier of learning processes.
The procedure is as follows:

1. Get a new data x_{ij}
2. Calculate the value of t in (3)
3. If t is greater than the value of t in t distribution with α or t is less than the value of minus t in t distribution with α, then the new data is detected.

Moreover, the unique feature of this method is that this test method represents various traditional statistical test method by changing the value of the hyper parameter n_0 as follows:

- When the value of the hyper parameter becomes enough large, then the method become equivalent to the Z test.
- When the value of the hyper parameter is equivalent to zero (is called Non information prior distribution), then the method is equivalent to the traditional t test.

Thus, the proposed method unifies various traditional test methods.

5. Online outlier detection system

We actually developed an e-learning platform system including this outlier detection system. The outlier detection system is shown in Figure 4. The system presents 1. learners names, 2. the learner's t value curve, and 3. the content with the irregular processes. We can know learners' learning processes on Online. detection system. The outlier detection system is shown in Figure 4. The system presents 1. learners names, 2. the learner's t value curve, and 3. the content with the irregular processes. We can know learners' learning processes on Online. If irregular processes are detected, the teacher investigates the learner's learning processes, and sends e-mail with some comments later.

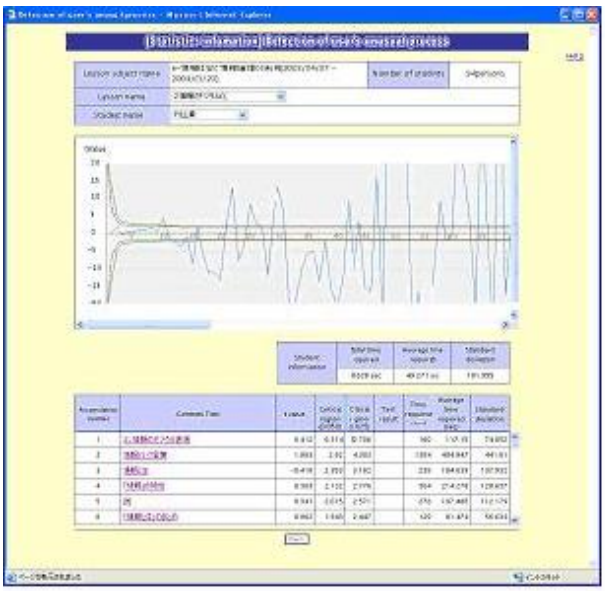


Figure 4. Online outlier detection system

6. Evaluation of the method by simulation experiments

Although the proposed method represents various statistical test methods by changing the value of the hyper parameter n_0 , it is not known how to determine value of the hyper parameter n_0. This section provides some

Table 1. Comparisons of the outlier detection methods by changing the hyper parameters

n	the probabilities which the methods incorrectly detect regular processes							the probabilities which the methods correctly detect irregular processes						
	Z test	n_0=0	n_0=1	n_0=5	n_0=10	n_0=15	n_0=20	Z test	n_0=0	n_0=1	n_0=5	n_0=10	n_0=15	n_0=20
0-10	.37	.07	.057	.10	.15	.21	.25	.89	.72	.74	.82	.86	.87	.88
10-20	.47	.12	.11	.19	.28	.36	.42	.98	.88	.90	.97	.99	.99	1.00
20-30	.44	.11	.10	.16	.24	.31	.37	.95	.84	.85	.92	.95	.95	.96
30-40	.47	.14	.12	.20	.28	.37	.43	.98	.91	.92	.97	.99	.99	.99
40-50	.46	.15	.14	.20	.29	.37	.43	.99	.97	.97	.99	.99	.99	1.00
50-60	.46	.15	.14	.21	.29	.37	.43	.99	.97	.98	.95	1.00	1.00	1.00
60-70	.46	.15	.14	.21	.29	.37	.43	.99	.98	.98	.99	1.00	1.00	1.00
70-80	.45	.15	.14	.21	.28	.37	.43	.99	.98	.99	.99	1.00	1.00	1.00
80-90	.44	.16	.15	.21	.29	.37	.43	.99	.99	.99	.99	1.00	1.00	1.00
90-100	.45	.16	.15	.22	.29	.37	.43	.99	.99	.99	.99	1.00	1.00	1.00

simulation experiments in order to determine the optimum value of the hyper parameter. The flow of the simulation experiment is as follows:

- Fix the learner j, and generate the random data from

$$x_{ij} = \frac{t_{ij} - \bar{t}_i}{s_i} = \mu_j + e_j$$

- Apply the proposed method to the generated data.
- These procedures are repeated in 1000 times.
- Calculates the probabilities which the methods correctly detect irregular processes and the probabilities which the methods incorrectly detect regular processes by changing the value of the hyper parameters.

The results are shown in Table 1. The column *n* in the table indicates the numbers of the random data sequences which are used for the outlier detection. For example, 0-10 indicates that the outlier detection procedure are completed using data from first data to 10^{th} data in the random data sequence. The calculated probabilities are given the average of the probabilities which the methods incorrectly detect regular processes and the probabilities which the methods correctly detect irregular processes by changing the value of the hyper parameter. The results show that the probabilities which the methods correctly detect irregular processes in each value of n become large when the value of the hyper parameter becomes large. When the value of the hyper parameter becomes large, the probabilities corresponding to each value of the hyper parameter become close to the probabilities in the Z test. On the other hands, the results also show that the probabilities which the methods incorrectly detect regular processes in each value of n become large when the value of the hyper parameter becomes small. In this case, the probabilities in the case of n_0=0 is equivalent to the probabilities in T test. Thus, we have to decide the value of the hyper parameter in consideration of the balance between two probabilities. This paper thinks the probabilities which the methods incorrectly detect regular processes important, we employ the value of the hyper parameter n_0=1 which minimizes the probabilities which the methods incorrectly detect regular processes.

week. The teachers instructed the students that they could study the contents of e-learning between Monday and Thursday, and they got e-mail with advices from the teachers. Therefore, the students have to learn the contents until Friday.

The figure 5 shows the learner 4's (outlier) detection curve corresponding to Figure 3. Parallel four lines in Figure 5 indicate the outlier detection line. For example, if the t value corresponding to a learning process exceed the top detection curve, it means that the learning process is irregularly too long. If the t value corresponding to a learning process exceed the bottom detection curve, it means that the learning process is irregularly too short.

In the case of Figure 5, the outlier processes appear in the contents 129-145. From Figure 3 and Figure 5, we should noted that the responses, which seems very long or very short comparatively in the Figure 3, are not always judged as the outlier processes. The reason is that the statistics value of t for the outlier detection is estimated with considering both the students ability for learning and the difficulty of each contents. For example, even if we find the responses which takes longer than the other responses, but we can not decide that the responses are outlier processes. Because, the contents may need longer time than the other contents. We can consider the case of very short response times. These features are quite different points from the traditional data mining methods using the outlier system , for example ,discovering robber crime from depositing money processes in the bank,

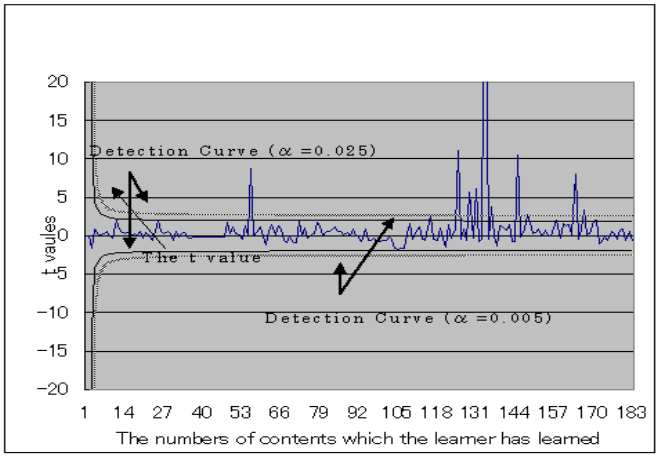
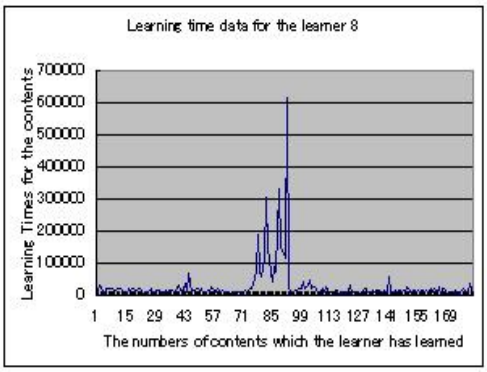
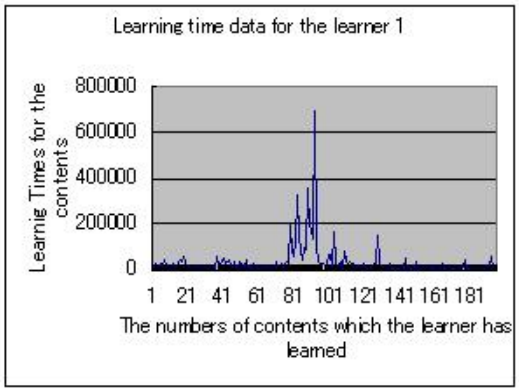
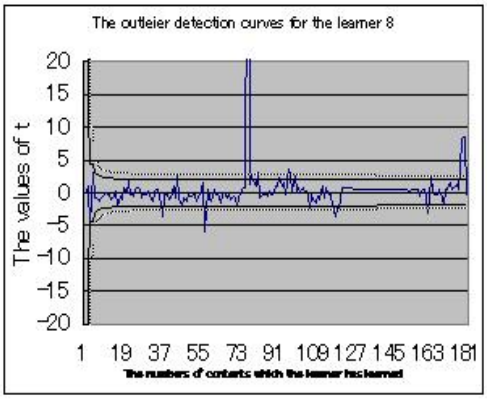


Figure 5. outlier detection curve corresponding to Figure 3







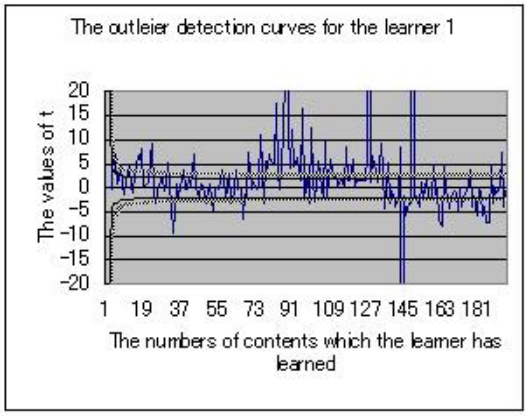


Figure 6. An example of detection curve in the case that there are few outlier processes

Figure 7. An example of detection curve in the case that there are many outlier processes

discovering an irregular invasion to computer network from using computer network processes, and so on.

The figure 6 shows an example of the raw learning time data and the detection curve in the case that there are few outlier processes. From the raw data, it seems that learning time for the contents 72-92 are irregularly long, however the learning processes for the contents 77 and 78 are detected as outliers, the other processes are not detected as outliers. Moreover, the figure 7 also shows an example of the law learning time data and the detection curve in the case that there are many outlier processes. It should be noted that the shape of the curve which shows the raw learning time data in Figure 7 is strongly similar to one in Figure 6. However, The bottom side of the figure 7 shows that there are many outliers in his learning processes. It is suspected that this learner must not have studied hard. Thus, the proposed method can detect outlier learning processes which we can hardly notice by

just analyzing the raw learning time data. In practical use for distance education, the teacher can not all students' learning processes log. Using this system, the teacher can detect some learners who should be taken care of, and send e-mail with adequate messages, for example, " Do you have any contents which you hardly understood?", "Are you studying hard ? ", " Are you bored ? " and so on. By this, students who have some problem in their learning can realizes that the teacher notice them, and then it is expected that it will derive their motivation for their learning.

7. Evaluation

This section provides some experiments to evaluate the effects of this system. The author gave the learners the following instruction. .

"If you feel that your learning process about this section has some problems, then please click the No-button. If you can understand it, then please click "Yes" button."

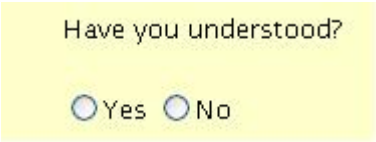


Figure 8. Yes-No Button used in the experiment

As the results, the learners' responses are divided into the following three responses: 1. Yes, 2. No, and no response. Here, we can consider that "Yes" responses have no problem, but "No" and no response might have some problems. It can be considered that "No" responses indicates that the learner can not understand the contents and the no response indicates that the learner might not have completed the contents.

Here, in the table 2, the conditional probabilities of the responses to the above question given the information about the learner's learning processes is detected or not as irregular processes.

Table2. Comparisons between detected irregular processes and learner's self statements

	When the learning process is detected as an irregular process	When the learning process is not detected as an irregular process
1."Yes "response	.24	.95
2."No"response	.31	.02
3. no response	..45	.03

The table 2 indicates that the detection system efficiently detect the learner's irregular processes. The probability of "Yes" response when the learning process is detected as an irregular process is comparatively large, but this indicates that this detection criterion is a little bit strict.

8. Conclusions

This paper proposed a method of online outlier detection of learners' irregular learning processes using the learners' response time data for the e-Learning contents. The unique features of this method are as follows: 1.It proposes an outlier detection method using Bayesian predictive distribution. 2. It is available for small sample, 3.It is convenient to calculate the predictive distribution. 4.On-line learning is realized on WWW. 5.It assists two ways instruction using data mining results for the learners' learning processes. 6. The outlier statistics is estimated by considering both the students' abilities and contents' difficulties. The system was utilized for actual classes, and then the results show the efficiency of the system.

The advantages of this outlier detection methods are 1. This method represents various statistical test methods by changing the value of hyper parameter, and 2. The model, which depends on a learner's learning ability and an item difficulty, provides the results that just seeing raw learning time data can not derive.

In this paper, we did not investigate why some students provide some irregular leaning processes. This is a future task.

References:

[1] Kawamura, S., "A note on Learner model by using access Log to WWW server", Educational Technology, Vol 98,No,188, pp.17-24, 1998

[2] Matsumoto,J., "Learning histories analysis system for educational assistance", Journal of Information Processing, Vol.40,No.9,1999, pp.3596-3607

[3] Matsui, T and Okamoto, T. " Discovery Science of Digital Portforio, Educational Technology , ET2000-88, pp.87-94,2000

[4]Barnett.V and Lewis.T, "Outliers in Statistical Data", John Willy & Sons, 1994

[5]Bonchi.F, Giannotti.F, Mainetto.G, and Pedeschi.D, A classification-based methodology for planning audit strategies in fraud detection", in Proc. Of KDD-99, pp.175-184, 1999

[6]Burge.P and Shaw-Taylor.J, "Detecting cellular fraud using adaptive prototypes, in Proc. of AI Approaches to Fraud Detection and Risk Management, pp.9-13, 1997

[7]Fawsett.T and Provost.F, Combining data mining and machine learning for effective fraud detection, in Proc. of AI Approachws to Fraud Detection and Risk Management, pp.14-19, 1997

[8]Lee.W, Stolfo.S.J, and Mok.K.W., "Mining in data-flow environment:experiment:experience in network intrusion detection, in Proc. of KDD99, pp114-124, 1999

[9]Ueno, M, "e-learning platform with Learning historial data and Dara mining", Proc. of ICALT 2002, in Kazan.

[10]Ueno,M: "LMS with irregular learning processes detection system", Proc. of E-learn2003, pp.2486-2493, 2003