Transductive Support Vector Machine Web Log Classifier for Identifying Potential Users

S. Chitra
Assistant Professor
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore - 641 018, Tamilnadu, India.
Email:chitra.sivakumar@ymail.com

B. Kalpana
Professor
Department of Computer Science
Avinashilingam Institute for Home Science
and Higher Education for Women
Coimbatore - 641 043, Tamilnadu, India.

Abstract--- Web usage mining deals with analyzing the log data of a web server. Web log analysis provides information to predict user behaviour and the efficiency of web site design. Users’ characteristics of a web site are analyzed by classifying the users into two categories: (1) Users with purchase interest, (2) Users without purchase interest. Prior to classification, the raw web logs are processed to identify the unique users and to construct sessions. This paper oversimplifies the method of classifying interesting users from a given set of web logs of an e-commerce web server. Modified Transductive Support Vector Machine (TSVM) algorithm is used to classify the web logs into these two categories.

Keywords---User identification, Session Construction, SVM (Support Vector Machine), Transductive SVM(TSVM), Web log Classification.

1. INTRODUCTION

The growth of World Wide Web is very rapid. Many research works are carried out to enhance the efficiency of services provided to the users over the internet. Web usage mining is an area where data mining techniques are applied to web server logs of a particular web site. It is also called web log mining. There are many types of web logs, but typically the log files share the same basic information such as client IP address, request time, requested URL, HTTP status code, referrer url, etc.

The interaction of any user with a website is recorded in the server’s log file. That may be used for various analysis and all the data mining techniques like association, classification, clustering, etc., may be applied on those log data. The usefulness lies in improving the performance of the web site in terms of design and content-building.

This paper concentrates on classifying the users of a web site into 2 categories as i) users who are really interested in buying one or more products that are displayed in the pages of the web site. ii) users who browse the pages of the site just to get familiarization about the site, i.e., visitors without purchase interest. This work is divided into three phases. In first phase the data cleaning and path completion processes are done. Second phase consist of session construction and user identification. Third phase focuses on web log classification. This paper uses semi-supervised Transductive SVM for web log classification. The remainder of the paper is organized as follows. In section 2, we discuss the related work. In section 3, Web Log Classification is discussed. In section 4 Transductive SVM Classification approach is discussed in detail. Results on the experiments conducted are discussed in section 5. Finally conclusion is given in section 6.

2. RELATED WORKS

Web log Classification using TSVM method is an approach that will improve the overall performance of the web server. Existing methods of weblog classification uses Decision Tree, Navie Bayesian Classification, etc.,.
Jie Zhang, *et. al.*, [1] proposed that the role of Web usage mining is very important for personalization of Web services. Several approaches have been proposed for extracting the required user sessions from the Web server logs.

Alka Gangrade, *et. al.*, [2] discussed about the techniques for privacy preserving classification under multi-party environment. Further, the two approaches, the classification model and secure multi-party computation algorithms have also been reviewed. The performance analysis of the algorithms has been concentrated in connection with the classification.

Mahesh Thylore Ramakrishna, *et. al.*, [3] use the data-centric view to refine the definition of Web mining. Data-centric view defines web mining with respect to the web data used whereas the alternative method process-centric view defines web mining as a serial collection of operations.

Classification algorithms discussed by Hanady Abdulsalam, *et. al.*, [4], have a set of training samples with labels, a set of test records without labels and classifier to label the test records and rerun the classifier.

Hidenao Abe [5] proposed a classification framework by combining the temporal pattern extraction and rule mining. This framework has been developed for mining if-then rules consisting of temporal patterns in left hand side of the rules. The right hand side helps us to predict both of important events and temporal patterns of important index.

Smith Tsang, *et. al.*, [6], discussed the problem of constructing decision tree classifiers on data with uncertain numerical attributes devised for decision tree construction. Lots of applications for this algorithm are also described in this paper.

C4.5 algorithm for data classification is discussed in Veronica S. Moertini [7]. The algorithm was experimented in utilizing C4.5 for varied dataset.

Enhanced C4.5 algorithm was introduced by Salvatore Ruggieri [8]. It improves C4.5 by adopting the best among the strategies for computing the information gain of continuous attributes. All the strategies adopt a binary search of the threshold in the whole training set starting from the local threshold computed at a node.

The Decision Tree’s can deal with one attribute per test node or with more than one. The former approach is called Univariate Decision Tree, and the latter is the Multivariate method. Thales Sehn Korting [9] explains the construction of Multivariate Decision Tree’s and the C4.5 algorithm, used to build such trees.

Mahdi Khosravi *et. al.*, [10] proposed a dynamic mining approach for modeling and predicting users’ navigation patterns. A.K. Santra, S. Jayasudha,[11] proposed a classification model using Naïve Bayesian classification method. They considered the page count, time spent on each page, number of pages visited, etc., as classifying attributes. The method discussed by Jeffrey Xu Yu, *et. al.*, [12] for classification motivated us to use TSVM classification model for web log data which takes less execution time and gives more accuracy.

### 3. Web Log Classification

Web usage mining will give us a way to analyze the user characteristics of a particular web site. The visitors of the site may be classified into two. 1. Users who are interested in buying the products posted in the web site. 2. Users who are not interested in the buying but browse the site by accident or they visit the site to familiarize the contents of the site. In addition to these two groups of customers, a special group may also be considered, “the network robots. Many search engines use network robots to scramble over the Web. The robots generate numerous access records in Web logs that seriously influence the discovery of customers’ patterns. In our method robots are cleaned before classification process starts. We have developed a classification approach that provide web site administrators to know the different type of visitors of their site. Classification algorithm uses the logs and manually create an attribute set for training phase[12].

#### 3.1. Web Site Visitor Characteristics

Extended log format contains ip-address, password, time stamp, url, status code, access methods, the transferred bytes, URLs of referrers’ pages, and user agents. We use some of these fields as classification attributes. Our Classification algorithm uses the attribute discretization as Jeffery Xu Zu, et al.,[12].

Browsers of a web site who are really interested in buying the products have the following characteristics:

- Visitors spend time to read the matter present in the page. They spend large amount of time to read the contents of the page. And also the time taken to
navigate from one page to another page is also large.

- Visitors read all the topics, and they search some specific topics also.
- They often use the HTTP POST mode (which sends data to a Web server and retrieves a response), because they’re interested in registering with Web sites and are willing to fill out forms with their own information.
- They often access images and graphic files.

On the other hand, visitors with no purchase interest exhibit these access patterns:

- They access many pages quickly to browse contents. The ratio between the time they need to read contents and the time they navigate from one page to another is almost 1.
- They don’t navigate down to low-level pages but rather access a large number of high-level child pages, because they’re not interested in any specific topics.
- They don’t often use POST mode, because they’re not interested in registering at Web sites.
- They don’t access images and graphic files.

On these bases, we classify two types of accesses (Visitors with purchase interest, and Visitors without purchase interest), we select eight attributes to construct our classifier.

Table 1 shows the eight attributes grouped into three types: temporal attributes (A1–A3), page attributes (A4–A7), and communication attributes (A8).

Discretization of the attribute values are given in table 2. Then, with customers’ browsing behaviour, we identify a small set of training data. The dataset’s labels should reflect the customers’ understanding of their own behaviors.

### 3.2 Classifier Design

Our algorithm classifies Web logs using Transductive Support Vector Machine Algorithm (TSVM). TSVM is used since our classification algorithms classify unlabelled samples also. TSVM uses a set of training data which have labels and a test data set that are to be labeled correctly using prediction. This is a semi-supervised classification algorithm where concrete training for labels cannot be given. Web logs don’t show whether a particular visitor of a web site is really interested in buying the product(s) of the web site.

### 3.3 Evaluation

To test this algorithm, we conducted experiments on e-commerce web sites’ data collected between Jan 2013 to July 2013 and discretize the data. Tables 1 and 2 show the classification attributes and discretized attribute values. The numbers of logs considered are 10,767. We selected 89 sessions for experimentation. We built the classifier using 16 records and the remaining 73 as test data. After data cleaning the count became 1,002. After user identification and session construction, the logs are classified. All the records have eight attributes. The number of the i\textsuperscript{th} element in a record indicates the i\textsuperscript{th} attribute value (table1). Table 3 shows a sample of 16 training records. For example, the first positive training record for a visitor with purchase interest (the first column of table 3) has eight values, each of which comes from table 2. The first value, zero, indicates that the value of attribute A1 is zero, which means that there is no night-accessing[12].

<table>
<thead>
<tr>
<th>Table 1. Classification Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Temporal Attributes</strong></td>
</tr>
<tr>
<td>A1 Accessing between midnight and 7 a.m.</td>
</tr>
<tr>
<td>A2 The total session time</td>
</tr>
<tr>
<td>A3 Statistics such as the time a visitor accesses the site, the total time a visitor stays at the site, and the different amounts of time a visitor stays on various pages</td>
</tr>
<tr>
<td><strong>Page Attributes</strong></td>
</tr>
<tr>
<td>A4 The total number of accessed pages during the whole session</td>
</tr>
<tr>
<td>A5 The accessing width (the number of child pages accessed from a single page)</td>
</tr>
<tr>
<td>A6 The accessing depth (the depth of the pages accessed from a single page)</td>
</tr>
<tr>
<td>A7 The percentage of graphic files requested compared to the total number of accessed pages</td>
</tr>
<tr>
<td><strong>Communication Attributes</strong></td>
</tr>
<tr>
<td>A8 Access methods (such as Get, POST, and Head) that visitors use to interact with the site</td>
</tr>
</tbody>
</table>
Table 2. Discretized Attribute Values

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value 0</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>No</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A2</td>
<td>≤ 2min</td>
<td>2-5min</td>
<td>5-15min</td>
<td>15-30min</td>
</tr>
<tr>
<td>A3</td>
<td>≤ 3 sec</td>
<td>3-30 sec</td>
<td>≥ 30 sec</td>
<td>-</td>
</tr>
<tr>
<td>A4</td>
<td>≤ 2 pages</td>
<td>2-5 pages</td>
<td>≥ 5 pages</td>
<td>-</td>
</tr>
<tr>
<td>A5</td>
<td>≤ 2 pages</td>
<td>2-5 pages</td>
<td>≥ 5 pages</td>
<td>-</td>
</tr>
<tr>
<td>A6</td>
<td>1 hierarchy</td>
<td>2-3 hierarchy</td>
<td>≥ 5 hierarchies</td>
<td>-</td>
</tr>
<tr>
<td>A7</td>
<td>0%</td>
<td>0-20%</td>
<td>20-50%</td>
<td>50-100%</td>
</tr>
<tr>
<td>A8</td>
<td>Use Get</td>
<td>Use POST</td>
<td>Use Head</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3. Sixteen Training Records Containing Data on Nine Attributes

<table>
<thead>
<tr>
<th>With purchase interest</th>
<th>Without purchase interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,0,1,2,2,1,3,0</td>
<td>0,0,0,0,0,0,0,0,0</td>
</tr>
<tr>
<td>0,0,0,2,2,3,0</td>
<td>0,2,1,0,1,3,0</td>
</tr>
<tr>
<td>0,0,2,1,1,1,0,1</td>
<td>0,0,0,0,0,0,3,0</td>
</tr>
<tr>
<td>0,3,2,2,2,0,1,0</td>
<td>0,0,1,1,1,1,0,0</td>
</tr>
<tr>
<td>1,0,0,0,0,0,0,0,1</td>
<td>0,0,0,1,1,1,0,0</td>
</tr>
<tr>
<td>0,2,1,2,2,1,3,0</td>
<td>0,3,1,0,0,0,0,0</td>
</tr>
<tr>
<td>1,2,2,1,1,0,2,0</td>
<td></td>
</tr>
<tr>
<td>1,2,2,2,2,0,1</td>
<td></td>
</tr>
<tr>
<td>1,1,2,2,2,1,1,0</td>
<td></td>
</tr>
<tr>
<td>1,1,2,1,1,1,0,1</td>
<td></td>
</tr>
</tbody>
</table>

4 TRANSDUCTIVE SVM CLASSIFICATION

4.1 Binary Classification

Given training data \((x_i, y_i)\) for \(i = 1 \ldots N\), with \(x_i \in Rd\) and \(y_i \in \{-1, 1\}\), learn a classifier \(f(x)\) such that

\[
f(x_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}
\]

i.e. \(y_i f(x_i) > 0\) for a correct classification.
4.2 Support Vector Machine

Training Sets and Prediction models
- Input/output sets X, Y
- Training set (x_1,y_1),..., (x_m,y_m)
- "Generalization": given a previously seen \( x \in X \), find a suitable \( y \in Y \).
  - i.e., want to learn classifier: \( y = f(x, \alpha) \), where \( \alpha \) are the parameters of the function.

For example, if we are choosing our model from the set of hyperplanes in \( \mathbb{R}^n \), then we have:
\[
f(x, \{w, b\}) = \text{sign}(w^T x + b).
\]

We can optimize the following:
\[
\min_{y_j, f \in F} C_1 \sum_{i=1}^{n_l} L(y_i f(x_i)) + C_2 \sum_{j=n_l+1}^{n} L \left( y_j f(x_j) \right) + J(f)
\]

Where \( f \) is a decision function in \( F \), a candidate function class, \( L(z) = (1 - z)_+ \) is the hinge loss, and \( J(f) \) is the inverse of the geometric separation margin.

In the linear case, \( f(x) = w^T x + b \) and \( J(f) = \frac{1}{2} ||w||^2 \). In the nonlinear kernel case, \( f(x) = (K(x, x_1), \ldots, K(x, x_n))^T w + b \), \( J(f) = \frac{1}{2} w^T Kw \), where \( K \) is a kernel satisfying Mercer's condition to assure \( w^T Kw \) with \( K = (K(x_i, x_j))_{i,j=1}^{n_l} \) being a proper norm.

4.4 Transductive SVM for Web Log Classification

Mercer’s condition to assure \( w^T Kw \) with \( K = (K(x_i, x_j))_{i,j=1}^{n_l} \) being a proper norm.

Input:
- training samples \((X_1, Y_1), \ldots, (X_m, Y_m)\)
- test samples \(X_i, \ldots, X_j\)

Output:
- predicted labels of the test samples \(X_i, \ldots, X_j\)
- return \((Y_i, \ldots, Y_j)\)

Step 1: (Initialization) Set initial value \( f^{(0)} \) as the solution of SVM with labeled data alone, and an precision tolerance level \( \epsilon > 0 \).

Step 2: At iteration \( k+1 \), solve yielding solution \( f^{(k+1)} \)

Step 3: (Stopping rule) Terminate when \( |s(f^{(k+1)}) - s(f^{(k)})| \leq \epsilon \). Then the estimate \( f \) is the best solution among \( f^{(k)} \); \( k=0, \ldots, 1 \)
Classification of Web logs using transductive SVM involves the following steps:

1. Identify the set of attributes, for this identification in this paper use nine attributes to classify the Web logs. These nine attributes are selected based on this visitor’s interest to purchase and non-interest. Classifier is constructed with the help of these nine attributes. Table 1 shows the nine attributes grouped into three types: temporal attributes (A1–A3), page attributes (A4–A7), and communication attributes (A8).

2. We discretize the attribute values; it is given in table 2.

3. Then identify the training data set with the help of customers. The data set’s label should reflect the customers’ understanding of their own behaviors.

4. Build a classifier using 16 training samples (with labels).

5. Predict the labels of the remaining 19 test records by using the transductive procedure.

6. Calculate precision, recall and accuracy.

7. Rerun the classifier by including the test records also.

5. Experimental Results

In the proposed method, modified transductive support vector machine algorithm is used for classifying an e-commerce website users into two categories, i.e., users with purchase interest and users without purchase interest. We are considering the logs of S&T Welcare Equipments (P) Ltd. (www.welcareindia.com) who do on-line business of fitness equipments. User’s log data is taken between Jan 2013 and July 2013. The numbers of logs considered are 10,767. We selected 89 sessions for experimentation. We built the classifier using 16 records and the remaining 73 as test data. After data cleaning the count became 1,002. After user identification and session construction, the logs are classified. Performance factors precision and recall are defined as follows:

\[ \text{Precision} = \frac{tp}{tp + fp} \]

\[ \text{Recall} = \frac{tp}{tp + fn} \]

Recall in this context is also referred to as the True Positive Rate or Sensitivity and precision is also referred to as Positive predictive value (PPV); other related measures used in classification include True Negative Rate and Accuracy. True Negative Rate is also called Specificity.

\[ \text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \]

Table 5. Comparison of two classes for customers after classifying sixteen training log records.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer with purchase interest</td>
<td>91</td>
<td>86</td>
</tr>
<tr>
<td>Customer with non purchase interest</td>
<td>78</td>
<td>70</td>
</tr>
</tbody>
</table>
Table 6 Accuracy comparison between TSVM and Decision Tree Algorithm

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer with purchase interest</td>
<td></td>
</tr>
<tr>
<td>Proposed method of TSVM classifier</td>
<td>97</td>
</tr>
<tr>
<td>Existing method of Decision tree algorithm</td>
<td>78</td>
</tr>
<tr>
<td>Customer with non purchase interest</td>
<td>85</td>
</tr>
</tbody>
</table>

Figure 6 Accuracy Comparison

From the above figure, it is clear the TSVM method performs better than Decision Tree method in terms of accuracy.

Table 7 Execution time comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Execution time in sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method of TSVM classifier</td>
<td>1.2</td>
</tr>
<tr>
<td>Existing method of decision tree algorithm</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 7 provides the execution time for TSVM classifier and decision tree algorithm. The results show that the TSVM method outperforms the Decision Tree algorithm in terms of execution time.

6. CONCLUSION

The performance of the proposed transductive SVM is compared with Decision Tree algorithm. Results show that the proposed algorithm works better than its Decision Tree counterpart with respect to execution time and accuracy. Moreover, the algorithm can be used in situations where the exact labels are not known for some training samples, i.e., applications that need semi-supervised learning can use this method for predicting the correct class.

REFERENCES


[9] Thales Sehn Korting, “C4.5 algorithm and


AUTHORS PROFILE

Mrs. S. Chitra is an Assistant Professor of Computer Science in Government Arts College, Coimbatore, Tamilnadu, India. She received her Masters’ degree in Computer Science from Avinashilingam University, Coimbatore. She has around 16 years of teaching experience at the post graduate and under graduate levels. Presently she is a Ph. D research scholar in Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore, Tamilnadu, India. Her areas of interest are Data Mining and Web Mining. She is a life member of The Indian Science Congress Association, Kolkata.

Dr. B. Kalpana is a Professor of Computer Science in Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore, Tamilnadu, India. She received her PhD in Computer Science from Avinashilingam University, Coimbatore. She specializes in Data mining. She has around 22 years of teaching experience at the post graduate and under graduate levels. She has published and presented papers in several refereed international journals and conferences. She is a member of the International Association of Engineers and Computer Scientists, Hongkong, Indian Association for Research in Computing Sciences (IARCS) and the Computer Society of India.