International Conference on Computational Science, ICCS 2013

The Application of Multiple Criteria Linear Programming in Advertisement Clicking Events Prediction

Fang Wang\textsuperscript{a}, Peng Zhang\textsuperscript{b}, Yanmin Shang\textsuperscript{b}, Yong Shi \textsuperscript{a,c,*}

\textsuperscript{a}Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
\textsuperscript{b}Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
\textsuperscript{c} College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA

Abstract

In advertisement industry, it is important to predict potentially profitable users who will click target ads (i.e., Behavioral Targeting). The task selects the potential users that are likely to click the ads by analyzing user's clicking/web browsing information and displaying the most relevant ads to them. In this paper, we present a Multiple Criteria Linear Programming (MCLP) prediction model as the solution. The experiment datasets are provided by a leading Internet company in China, and can be downloaded from track2 of the KDD Cup 2012 datasets. In this paper, Support Vector Machines (SVM), Logistic Regression (LR), Radial Basis Function Network (RBF Network), k-Nearest Neighbour algorithm (KNN) and NaïveBayes are used as five benchmark models for comparison. The results indicate that MCLP is a promising model in behavioral targeting tasks.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license. Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science

Keywords: Behavioral Targeting; Multiple Criteria Linear Programming (MCLP); Support Vector Machines (SVM); Logistic Regression (LR); Radial Basis Function Network (RBF Network); k-Nearest Neighbour algorithm (KNN); NaïveBayes;

1. Introduction

Here introduce the paper. With the increasing of Internet users, online advertising becomes an important advertising market and provides a major source of advertising revenues [1]. For web-based businesses, Internet advertising has become a major source of revenue. Internet advertising revenues in the U.S. reached $9.26 billion for the third quarter of 2012, making the quarter the biggest on record, according to the latest IAB Internet Advertising Revenue Report figures released by the Interactive Advertising Bureau (IAB) and PwC US
Major online publishers such as Yahoo!, Microsoft and Google have enthusiastically embraced this business model.

The commercial value of advertisement on the Web depends on whether users click on the advertisement. The advertisements click has a significant impact on the Internet industry. It allows Internet companies to identify most relevant ads for each user and improve user experiences. Internet Behavioural targeting (BT) leverages user’s online activities to select the ads most relevant to users to display, which is a promising technique to improve the efficiency of online advertising.

There has been a lot of research in Behavioural Targeting. A well-grounded statistical model of BT predicts click-through rate (CTR) of an ad from user behaviour, such as ad clicks and views, page views, search queries etc. The CTR is used in search advertising to rank ads and price clicks.

In this paper we view this task as a binary classification problem and address it utilizing Multiple Criteria Linear Programming (MCLP) [3] and compare it with other five well-known classification methods. The results of the experiment demonstrate that MCLP is a good method in the research field of Behavioral Targeting. The datasets [4] used for testing comes from track2 of the KDD Cup 2012. A major challenge is to create efficient features. Feature creation is one of the most important steps in solving a supervised learning problem. We compared different methods and then chose two of them to create the features.

The paper is structured as follows: Section 2 reviews related work. Section 3 describes our behaviour data. Section 4 introduces a Two-Class MCLP Data Mining Model and its Algorithm. Section 5 is the experiment. We conclude the paper in Section 6 with future possible work.

2. Related Work

Much attention has been paid on the advertisement research recently. The best way to maximize the commercial value of advertisements is to display the ads to people who are interested in it. However, there are some issues to be dealt with, such as matching relevant advertisements for a query, ranking of the candidate advertisements, deciding how to display the advertisements on the search result page, click prediction and analysis for the presenting advertisements and pricing of the advertisements. Several machine learning algorithms such as logistic regression, linear Poisson Regression, Online Bayesian Probit Regression, support vector machines (SVM) [5,6,7,8,9] and Latent Factor Model have been adopted to predict the clicks of advertisements presented for a query. Since the size of online data is usually huge, online data stream classification analysis can be very helpful in Behavioral Targeting field [10,11,12].

Behavioral Targeting contains three pricing models, which are Pay-Per-Click (PPC), Pay-Per-Impression (PPI) and Pay-Per-Transaction (PPT). The popular one is PPC. For the PPC model, both the advertiser and the search engine companies wish users to click the advertisements. Therefore, Behavioral Targeting is a good way to solve this problem because it reduce advertiser’s cost and increases search engine companies’ profit simultaneously. In this paper, we exploit binary classification models to solve the problem. We classify training instances into two groups, in section 3, the classification of data will be covered in detail (We don't simply categorize those users who click the advertisements as positive samples and who not click the advertisements as negative samples). Then we train a classifier to discriminate positive samples from negative ones. Naïve Bayes, logistic regression and SVM have been used in Behavioral Targeting. In this paper, we explore a new model here: MCLP and compare it with other five famous classification methods.

Multiple Criteria Linear Programming(MCLP) is a promising optimization-based classification model [13, 14] and has extended to a family toolbox [15]. MCLP has many successful applications including credit card portfolio management [16], credit card risk analysis [17], firm bankruptcy prediction [18, 19], network intrusion detection [20, 21], medical diagnosis and prognosis [22] and classification of HIV-1 mediated neuronal dendritic and synaptic damage [23].

3. Feature Creation & Selection
In this paper, the training sample comes from track 2 of the KDD Cup 2012 datasets. The training set contains 155,750,158 instances that are derived from log message of search sessions, where a search session refers to an interaction between a user and the search engine. During each session, the user can be impressed with multiple ads, then, the same ads under the same setting (such as position, depth) from multiple sessions are aggregated to make an instance in the datasets. Each instances can be viewed as a vector (#click, #impression, DisplayURL, AdID, AdvertiserID, Depth, Position, QueryID, KeywordID, TitleID, DescriptionID, UserID). It means that under a specific setting, the user (UserID) has been impressed with the ad (AdID) for #impression times, and has clicked #click times of those. In addition to the instances, the datasets also contains token lists of query, keyword, title and description, where a token is a word represented by its hash value. The gender and segmented age information of each user are also provided in the datasets. The test set contains 20,297,594 instances and shares the same format as the training set, except for the lack of #click and #impression. The test set is generated with log messages that come from sessions latter than those of the training set. Detailed information about the datasets of KDD Cup 2012 can be founded in [4].

Feature creation and selection are a major challenge in this paper. We use two different training sets with different feature creation and selection methods in this paper, which are called T-Set-1 and T-Set-2, respectively.

3.1. Feature creation method for T-Set-1

In T-Set-1, the bag of words model was used. This method is frequency-based method that is used to predict the probability of each presented word on a clicked instance based on each feature (tokens). Then, we built the whole feature space by combining the query dictionary and ad dictionary.

3.2. Feature creation method for T-Set-2

Two kinds of features, original features and synthetic features, were used for modeling in this method.

(1) Original Features: The original feature set contains discrete features and continuous features. The discrete features are the unique ID of each ad, advertiser, query, keyword, tile, description, token, gender and age for one user, depth and position of ads, and the displayed URL. The continuous features are the click-through rates of each value of the discrete features. When a discrete feature is being used, the corresponding click-through rate will be activated and adopted as a continuous feature.

(2) Synthetic Feature: First of all, we join any two original discrete features with each other and use them as synthetic features. We also test some 3-tuple features but only the QueryID_AdID_UserID is available. Since most 3-tuple features are too sparse and seldom activated. Secondly, we join the original discrete features with each of the tokens. Position information is added to the original discrete features to generate one 2-tuple position-based feature. Bigram features are also adopted for analyzing the queries, titles and descriptions.

3.3. Categorization method for positive/negative samples

We analysed the dataset comprehensively and think in-depth for predicting accurately. Let's consider:

Advertisement 1: The time of display is 10, the time of click is 0.
Advertisement 2: The time of display is 10, the time of click is 1.
Advertisement 3: The time of display is 10, the time of click is 8.

From above, we can see that the gap between advertisements 2 and 3 is bigger than the gap between advertisements 1 and 2. If we simply categorize those samples based on click times, those with click times greater than 1 are categorized as positive samples and those less than 1 as negative samples, then the advertisement 2 and 3 are both labeled as 1 while the label of advertisement 1 is -1. In this situation, the
influence of advertisement 2 and 3 are the same. However, as the time of click 0 and 1 is closer than 1 and 8, it is not reasonable. Therefore, we treat the click-through-rate as a probability problem. For one wonderful advertisement, someone will click it while others won’t. Therefore, in this paper, we calculate the click-through-rate (CTR) of each instance, and the average CTR. Then we compared each instance's CTR with the average CTR, if it is greater than the average CTR, the label of the instance should be 1, otherwise, it should be 0. The formula to calculate the CTR is described as below:

\[
\text{Click -Through -Rate (CTR)} = \frac{\# \text{click} + \alpha \times \beta}{\# \text{impression} + \beta}
\]

where \( \alpha = 0.05, \beta = 75 \), that obtained from the experiment.

3.4. Normalization

Since the ranges of all the variables’ value are significantly different, a linear scaling transformation needs to be performed for each variable. The transformation expresses as below:

\[
x_n = \frac{x_i - \min(x_i, K, x_n)}{\max(x_i, K, x_n) - \min(x_i, K, x_n)}
\]

where \( x_n \) is the normalized value and \( x_i \) is the instance value.

4. A Two-Class MCLP Data Mining Model and Its Algorithm

Using MCLP [24,25], we can optimize maximizing the minimum distances (MMD) and minimizing the sum of the deviations (MSD) simultaneously, producing better data separation than by linear discriminate analysis. According to the concept of Pareto optimality, we can seek the best trade-off of the two measurements [15, 17]. In this section, we outline the structure of a two-class MCLP model.

Given any two predefined classes \{G: Good and B: Bad\} for a datasets. Given training samples \( T_n = \{A_1, A_2, ..., A_n\} \), where \( n \) is the total number of records in the training samples. Each training instance \( A_i \) has \( r \) attributes. This data mining model is used to determine the coefficients for an appropriate subset of the variables, denoted by \( X = (x_{1:i}, x_j) \), and a boundary value \( b \) to separate two classes: G (Good) and B (Bad) with minimizing the overlapping; that is, if \( A_i X < b, A_i \in G \) and if \( A_i X > b, A_i \in B \), where \( A_i \) is the vector value of the subset of variables from the datasets and the symbol “ \( \in \) ” means “belongs to” . Note that when \( A_i X = b \), \( A_i \) belongs to either G or B. The geometric meaning of the model is shown in Fig.1. (a).

To measure the separation of G and B, we define:
- \( \alpha_i \) = the overlapping of a two-class boundary for case \( A_i \) (external measurement);
- \( \beta_i \) = the distance of case \( A_i \) from its adjusted boundary (internal measurement);

We use ( ) to represent \( A_i \) of Good and (•) to represent \( A_i \) of Bad. Fig.1. (a) Shows that our goal is to minimize the sum of \( \alpha_i \) and maximize the sum of \( \beta_i \) simultaneously. As a result, two groups of data represented in Fig. 1. (a) will be pulled away. Therefore, this model can be written as:

\[
\text{Minimize} \sum \alpha_i\text{ and Maximize} \sum \beta_i
\]

Subject to:
\[ A_i X = b + \alpha_i - \beta_i, A_i \in G, \]
\[ A_i X = b - \alpha_i + \beta_i, A_i \in B, \]

where \( A_i \) are given, \( X \) and \( b \) are unrestricted, and \( \alpha_i \) and \( \beta_i \geq 0 \).

\[ \text{Fig.1. (a) geometric meaning of MCLP; (b) geometric meaning of compromise solution of MCLP} \]

To facilitate the computation, the compromise solution approach [26] can be employed to reform the above model so that we can systematically identify the best trade-off between \(-\sum \alpha_i\) and \(\sum \beta_i\) for an optimal solution. To explain, we assume the “ideal value” of \( -\sum \alpha_i \) be \( \alpha^* > 0 \) and the “ideal value” of \( \sum \beta_i \) be \( \beta^* > 0 \). Then, if \(-\sum \alpha_i > \alpha^*\), we define the regret measure as \( d_a^+ = -\sum \alpha_i - \alpha^* \); otherwise, it is 0. If \(-\sum \alpha_i < \alpha^*\), the regret measure is defined as \( d_a^- = \alpha^* + \sum \alpha_i \); otherwise, it is 0. Similarly, we define regret measure \( d_b^+ = \sum \beta_i - \beta^* \) when \( \sum \beta_i > \beta^* \) and \( d_b^- = 0 \) otherwise; regret measure \( d_b^- = \beta^* - \sum \beta_i \) when \( \sum \beta_i < \beta^* \) and \( d_b^- = 0 \) otherwise. Thus, we have (1) \( \alpha^* + \sum \alpha_i = d_a^- - d_a^+ \) (2) \( \alpha^* + \sum \alpha_i = d_a^- - d_a^+ \) and (3) \( d_a^- + d_a^+ \geq 0, \beta^* - \sum \beta_i = d_a^- - d_a^+ \), |\( \beta^* - \sum \beta_i \)| = \( d_a^- + d_a^+ \), and \( d_b^- + d_b^+ \geq 0 \). The two-class MCLP model has evolved to the following model:

\[ \text{Minimize } d_a^- + d_a^+ + d_b^- + d_b^+ \]
\[ \text{Subject to: } \alpha^* + \sum_i \alpha_i = d_a^- - d_a^+, \]
\[ \beta^* - \sum_i \beta_i = d_a^- - d_a^+, \]
\[ A_i X = b + \alpha_i - \beta_i, A_i \in G, \]
\[ A_i X = b - \alpha_i + \beta_i, A_i \in B, \]
\[ \alpha_i, \beta_i, d_a^-, d_a^+, d_b^-, d_b^+ \geq 0 \]

where \( A_i, \alpha^*, \text{and } \beta^* \) are given, \( X \) and \( b \) are unrestricted, and \( \alpha_i, \beta_i, d_a^-, d_a^+, d_b^-, d_b^+ \geq 0 \). The geometric meaning of compromise solution of MCLP is shown in Fig.1. (b).

5. Experiment

In this paper, we used two files (Dataset1, Dataset2) for training, which were generated and selected by the methods mentioned in section 3, respectively. Dataset1 was the subset of the T-Set-1, Dataset2 was the subset...
of the T-Set-2 and Principal Component Analysis (PCA) was then used for feature selection. We use 10-fold cross validation for both of the two training sets. Both Dataset1 and Dataset2 contained exactly 6,000 records for modeling. The amounts of positive samples and negative samples are equal. The baseline models including Support Vector Machines (SVM), Naïve Bayes, Logistic Regression (LR), Radial Basis Function Network (RBF Network) and k-Nearest Neighbour algorithm (KNN).

5.1. Confusion Matrix

In this paper, we use Confusion Matrix for the performance analysis:
TP (True Positive) = The number of records in the first class that has been classified correctly;
FP (False Positive) = The number of records in the second class that has been classified into the first class;
TN (True Negative) = The number of records in the second class that has been classified correctly;
FN (False Negative) = The number of records in the first class that has been classified into the second class.
Then we have four different performance measures:

\[
\text{Specificity} = \frac{TN}{TN + FP}; \\
\text{Sensitivity} = \frac{TP}{TP + FN}; \\
\text{False Positive Rate} = \frac{FP}{TN + FP}; \\
\text{False Negative Rate} = \frac{TN}{FN + TN}.
\]

5.2. ROC Graphs

Receiver Operating Characteristics (ROC) graph is a useful technique for organizing classifiers and visualizing their performance. ROC graphs are commonly used in medical decision making, and in recent years have been increasingly adopted in the machine learning and data mining research communities. In addition to being a generally useful performance graphing method, they have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs. These characteristics have become increasingly important as research continues into the areas of cost-sensitive learning and learning in the presence of unbalanced classes.

5.3. Result

As shown in table 1 and 2, we can see that the results of the six models agree with each other quite well, the results trained with Dataset1 and 2 are closer too and both results are in reasonable agreement. From table 1 and 2 we can see that the results of the evaluation criteria sensitivity and specificity are very closer with using the model Multiple Criteria Linear Programming (MCLP), which means that the stability of MCLP is better than other models. The performance of MCLP is relatively better among all the models. The results of the experiment demonstrate that MCLP is a good model in the research field of Behavioral Targeting, comparing with the traditional mathematical tools in classification, such as neural networks, decision tree, and statistics, MCLP is simple and direct, free of the statistical assumptions, and flexible by allowing decision makers to play an active part in the analysis.

We managed to obtain the best parameters of MCLP and LibSVM by running our homemade programs.
Table 1. Comparisons among different models of Dataset1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification Result</th>
<th>Best Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Specificity</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>MCLP</td>
<td>0.951388</td>
<td>0.943843</td>
</tr>
<tr>
<td>SVM</td>
<td>0.936</td>
<td>0.973</td>
</tr>
<tr>
<td>LR</td>
<td>0.956</td>
<td>0.924</td>
</tr>
<tr>
<td>RBF Network</td>
<td>0.942</td>
<td>0.85</td>
</tr>
<tr>
<td>KNN</td>
<td>0.957</td>
<td>0.951</td>
</tr>
<tr>
<td>NaïveBayes</td>
<td>0.929</td>
<td>0.851</td>
</tr>
</tbody>
</table>

Table 2. Comparisons among different models of Dataset2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification Result</th>
<th>Best Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Specificity</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>MCLP</td>
<td>0.953902</td>
<td>0.953035</td>
</tr>
<tr>
<td>SVM</td>
<td>0.932</td>
<td>0.971</td>
</tr>
<tr>
<td>LR</td>
<td>0.965</td>
<td>0.931</td>
</tr>
<tr>
<td>RBF Network</td>
<td>0.941</td>
<td>0.861</td>
</tr>
<tr>
<td>KNN</td>
<td>0.946</td>
<td>0.959</td>
</tr>
<tr>
<td>NaïveBayes</td>
<td>0.93</td>
<td>0.875</td>
</tr>
</tbody>
</table>

Figure 2 (a) and (b) showed the ROC curves of NaïveBayes (×), Logistic (+), RBFNetwork (◊), IBk (△) and LibSVM (▼) on Dataset1 for positive samples and negative samples, respectively. Figure 3 (a) and (b) showed the ROC curves of Logistic (×), RBFNetwork (+), NaïveBayes (◊), IBk (△) and LibSVM (▼) on Dataset2 for positive samples and negative samples, respectively. Figure 4 showed the ROC curves of MCLP
on Dataset1 and 2 respectively. In Figure 2 and 3 (a), the horizontal axis represents the False Positive rate and the vertical axis represents the True Positive rate. In Figure 2 and 3 (b), the horizontal axis represents the False Negative rate and the vertical axis represents the True Negative rate. In Figure 4, the horizontal axis represents the False Positive rate and the vertical axis represents the True Positive rate. For the curves, the more close to the upper left corner is the better one, which means the True Positive rate is higher. Usually, the diagonal was used as the baseline. The ROC curves should be above the diagonal. The ROC curves in the figures demonstrate that our experiment results are meaningful in Behavioral Targeting field. The biggest contribution of this paper is that a new model (Multiple Criteria Linear Programming) was proposed and proved to be useful and valuable in Behavioral Targeting field.

Fig.2. (a) Comparisons of ROC curves among different models on Dataset1 for positive samples; (b) Comparisons of ROC curves among different models on Dataset1 for negative samples.

Fig.3. (a) Comparisons of ROC curves among different models on Dataset2 for positive samples; (b) Comparisons of ROC curves among different models on Dataset2 for negative samples.
6. Conclusions

In this paper, we regard advertisement clicking events as a binary classification problem, and present a multiple criteria linear programming (MCLP) algorithm to predict if a user will click an advertisement or not. Applying MCLP to deal with advertisement clicking events is new in Behavioral Targeting field. The experiment results demonstrate that MCLP is an efficient method in predicting advertisement clicking events. In the future, on the one hand, we will extend the method to a three-group classification; on the other hand, we will extend the method by integrating with other models (ensemble model) to improve the prediction result. Due to many potential customers exist on the Internet, user’s social data will be added into the training sample to solve the data sparse problem.

Acknowledgements

This work has been partially supported by grants from National Natural Science Foundation of China (Nos.70921061, 11271361, 71201143, 61003167), the CAS/SAFEA International Partnership Program for Creative Research Teams, Major International (Ragional) Joint Research Project (No.71110107026), the President Fund of GUCAS.

References