Research Article

Application of Decision Tree-Based Classification Algorithm on Content Marketing

Yi Liu1,2 and Shuo Yang3

1School of Business, Macau University of Science and Technology, Cotai, Macao, China
2School of Business, Guangdong Polytechnic of Science and Technology, Zhuhai, Guangdong, China
3School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, Guangdong, China

Correspondence should be addressed to Yi Liu; 1709853gbm30014@student.must.edu.mo

Received 6 January 2022; Revised 26 January 2022; Accepted 27 January 2022; Published 4 March 2022

Academic Editor: Miaochoa Chen

Copyright © 2022 Yi Liu and Shuo Yang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traditional content marketing methods resort grossly to market requirements but barely obtain relatively accurate marketing prediction under loads of requirements. Machine learning-based approaches nowadays are widely used in multiple fields as they involve a training process to deal with big data problems. In this paper, decision tree-based methods are introduced to the field of content marketing, and decision tree-based methods intrinsically follow the process of human decision making. Specifically, this paper considers a well-known method, called C4.5, which can deal well with continuous values. Based on four validation metrics, experimental results obtained from several machine learning-based methods indicate that the C4.5-based decision tree method has the ability to handle the content marketing dataset. The results show that the decision tree-based method can provide reasonable and accurate suggestions for content marketing.

1. Introduction

Content marketing is normally regarded as a management process responsible for identifying, anticipating, and satisfying customer requirements profitably in the context of content via electronic channels [1]. The designation of research methods in content marketing has received plenty of attention for more than 100 years. Traditionally, content marketing strategies are determined based on different developments technologies, market requirements and expectations, growth in knowledge, and so on [2]. Machine learning-based methods have received more attention in the past several decades because they have the ability to analyze historical data and predicting future potential behaviors and activities more effectively [3, 4]. When there exist newly available data, managers can capture relatively accurate estimations from machine learning prediction models to tackle underlying challenges, such as missing outcomes and stimulus sampling [5]. Due to the fast development of machine learning, the current content marketing strategies tend to be digital-driven, meaning that companies put more attention to the consumer concentration on the Internet or historical habitats [6]. In this regard, the data scale and dimensions become huge, meaning that the traditional marketing methods are not suitable. The major motivation of this paper is therefore to investigate the potential of machine learning methods in dealing with the high dimensionality of attributes of marketing data.

Based on different data requirements, machine learning methods can be divided into three parts [7]. The first part refers to supervised learning-based methods that require an independent training process and predict testing data using an already learned model. The representative supervised models include sparse/collaborative representation [8–10], support vector machine [11–13], ensemble learning [14–17], and so on. The second part refers to unsupervised learning-based models that do not demand training samples and determine entire classes by considering the correlations among samples. The unsupervised models are also called clustering methods, such as K-means [18, 19], ISODATA
Among the above-mentioned machine learning methods, decision trees are the most discussed scalable multivariate methods in machine learning, which intrinsically follow the process of human decision making. Unlike emerging deep learning-based methods \([27–31]\) that require strict support, decision trees consider the latent structure of training data. That is to say, they split the training samples into bins and each of them is verified after calculating the most prevalent split variable by means of a specific metric, such as information gain value, Gini index, and entropy metrics. Besides, decision tree-based methods show at least three advantages. First, the decision tree model to some degree is easy to follow and implement, meaning that customers do not have to know complex and latent background knowledge in the training procedures. Second, the decision tree-based learning models do not pay sufficient attention to the preparation process of samples as it to some degree is useless and time consuming. The main reason behind this fact is that the decision tree-based model equally treats the attributes of samples as the learning foundations, which makes it possible to handle a relatively huge data scale. Third, the process of verifying effectiveness and robustness of the decision tree model can be easily measured. Figure 1 visually illustrates the processing of classification using a decision tree.

Chou \([32]\) proposed a classification and regression tree method in 1984, called CART. This method is a representative nonparametric learning strategy that produces classification and decision trees based on the status of dependent variables. In 1986, Quinlan proposed an iterative dichotomiser 3 method, called ID3 \([33]\). In this method, information gain is considered to determine the potential of the next nodes, meaning that the nodes with high information gain will be split. Xiaoliang et al. \([34]\) extended Quinlan’s work and developed a well-known C4.5 method in 2009. Similar to ID3, C4.5 uses a well-known information gain metric. Besides, it also deals well with continuous inputs because it uses thresholds and splits attributes when the values are higher than the threshold. In 2012, Patil et al. \([35]\) further extended C4.5 and ID3 and proposed C5.0 classification method. C5.0 not only has low time and memory consumptions compared to C4.5 and ID3 but also is efficient as it splits the nodes in terms of the field that has the maximum information gain. Recently, publications related to decision tree-based machine learning methods exist in different variants such as random forest \([36, 37]\), gradient boosting decision tree \([38]\), regression decision tree \([39, 40]\), and so on.

Among loads of decision tree-based methods, this paper specifically considers C4.5 and applies it to content marketing data analysis. This method is widely used because its classifying rules follow the human’s thinking process and it also has the ability to produce desirable results. Based on four validation metrics, we conduct several important experiments on the bank content marketing dataset for the purpose of verifying the performances of C4.5 and other comparison methods. The main contribution of this paper is that we introduce a decision tree-based method to the content marketing field to efficiently improve the marketing capability.

2. Materials and Methods

2.1. Validation Metrics. Validation metrics including entropy, information gain, Gini index, gain ratio, and so on are the most important standards to determine whether the current node should be split. Entropy verifies the uncertainties among a set of random variables. When random variables have high entropy values, it means that the uncertainty of the random variables is equally high and the current should be further divided into two nodes. The entropy information of such random variables is given as follows:

$$H(X) = - \sum_{x \in X} p(x) \log p(x),$$  \hspace{1cm} (1)

where \( \log \) denotes a base 2 logarithm and they have the same meaning in the subsequent formulas.

Information gain represents the difference between the entropy obtained before performing splitting and that after performing splitting. In this regard, the standard of splitting relies on the method that can produce maximum information gain as it means better classification performances. The information gain is

$$\text{gain}(X, A) = H(X) - H(X|A),$$  \hspace{1cm} (2)

where \( H(X|A) \) means the entropy after excluding \( A \) attributes from \( X \), given by
\[ H(X|A) = \sum_{a \in A} p(a) \times H(X|A = a) \]
\[ = \sum_{a \in A} p(a) \times \left[ -\sum_{x \in X} p(x|A = a) \log p(x|A = a) \right]. \]

(3) By calculating the entire information gain ratio under possibilities, we can obtain two optimal splitting subsets corresponding to maximal gain ratio results. We record the corresponding splitting threshold.

2.3. Pruning Process in C4.5. The establishment process of the decision tree depends heavily on training samples, meaning that the fitting effect on the given training data will occur. It is worth mentioning that, however, the decision tree-based learning model to some extent is complex for the testing data, triggering low classification accuracies, a.k.a. overfitting. Therefore, the process of simplifying the decision model is highly demanding, resulting in a crucial step, i.e., pruning.

C4.5 algorithm adopts the BEP (Bayesian error pruning) method. The BEP method was proposed by Quinlan, which is a top-down pruning method. It determines whether to prune the subtree according to the error rate before and after pruning, so it does not need a separate pruning dataset.

For a leaf node, assume that it involves \( N \) samples with \( E \) errors. Then, the error ratio is \( (E + p)/N \), where \( p \) denotes a penalty factor. Then, for a subtree with \( L \) leaf nodes, its error ratio is

\[ \text{error ratio} = \frac{\sum_{i=1}^{L} E_i + pL}{\sum_{i=1}^{L} N_i}, \]

where \( E_i \) denotes the sample numbers with regard to \( i \)-th leaf and \( N_i \) denotes the total number of samples in \( i \)-th leaf node.

Suppose a subtree misclassifies a sample with a value of \( 1 \) and correctly classifies a sample with a value of \( 0 \), then the number of misclassifications of the subtrees follows a Bernoulli distribution, so their statistical information, i.e., mean and standard deviation values, can be obtained by

\[ \text{error mean} = \text{error ratio} \times \sum_{i=1}^{L} N_i, \]

\[ \text{error Std} = \sqrt{\text{error ratio} \times \sum_{i=1}^{L} N_i (1 - \text{error ratio})}. \]

After replacing the subtree by leaf nodes, their error values can be determined by

\[ \text{error ratio}' = \frac{E' + p}{N'}, \]

where \( E' = \sum_{i=1}^{L} E_i \) and \( N' = \sum_{i=1}^{L} N_i \).

Note that the number of misclassifications of leaf nodes also follows a Bernoulli distribution. The mean values of the number of misclassifications of leaf nodes are defined as

\[ \text{error mean}' = \text{error ratio}' \times N'. \]

Here, pruning can be performed if it satisfies the following formula:
In order to visually describe the process of performing C4.5 decision tree on given data, this paper provides the corresponding flowchart of C4.5 (see Figure 2).

3. Result Analysis and Discussion

3.1. Experimental Parameters. The entire experiment involved in this paper was conducted on a PC with Intel Core i7-10700 processor at 2.90 GHz and 16 GB RAM. The GPU is NVIDIA GeForce GTX 1660 Super. The comparison algorithms in this paper involve nearest neighbors (the number of nearest neighbors is fixed at 3), linear SVM (the penalty parameter is fixed at 0.025), neural network (the regularization parameter is fixed at 1 and other parameters are fixed under default settings in sklearn package), Naive Bayes, and random forest (the maximum depth is fixed at 5).

3.2. Datasets. This paper considers a bank campaign content marketing dataset, which was originally derived from the Portuguese banking institution. This dataset has 50 user features by investigating multiple content marketing campaign activities based on traditional phone calls in order to find the real information that if users are interested in purchasing the products in the near future. Figure 3 visually describes the correlations among several features using a heatmap.

3.3. Validation Metrics. In order to accurately describe the experimental accuracy, this paper uses four standard evaluation metrics derived from the confusion matrix for evaluation purposes including accuracy, precision, recall, and F1 score. The above four evaluation criteria are calculated based on the confusion matrix. Figure 4 gives the visual explanations of the confusion matrix. For the classification results of a given binary classification model, true positive (TP) represents an outcome where the model correctly predicts the positive class; false positive (FP) represents the results that indicate a given condition exists when it does not; true negative (TN) is an outcome where the model correctly predicts the negative class; false negative (FN) is an outcome where the model incorrectly predicts the negative class. Based on four definitions, accuracy (acc), precision (prec), recall, and F1 score (F1) can be further defined by

\[
\text{acc} = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
\text{prec} = \frac{TP}{TP + FP},
\]

\[
\text{recall} = \frac{TP}{TP + FN},
\]

\[
F1 = \frac{2 \times TP}{2 \times TP + FP + FN}.
\]

3.4. Experimental Results. Figure 5 displays the resulting trend of the decision method when varying the max depth from 1 to 15. As can be seen from Figures 5(a) and 5(b), when the number of max depths is fixed as 4, the decision tree method shows the best training accuracies in Gini index and gain ratio metrics. Normally, the max depth of decision tree means the maximum depth to which a tree is allowed to grow. However, a deeper tree will trigger a complex model. Specifically, for the training stage, if we increase max depth, the training error will always decrease (or at least not increase). For the testing stage, if depth is set too high, the decision tree may result in overfitting problems, which increase test errors. On the contrary, if it is set too low, the decision tree barely captures latent information from

\[
\text{error mean} + \text{error Std} \geq \text{error ratio}'.
\]
training data. Hence, in this paper, we fix the number of max depths of the decision tree as 4.

Figure 6 visually illustrates the node construction process of the decision tree. Owing to the fact that the number of useful features is too large, the max depth of the decision tree is fixed at 3 for better visualization. As can be seen from Figure 4, the splitting manner highly depends on a determined metric and a threshold. When the Gini index of a node is higher than a threshold, then it will be split into two new nodes.

Generally, each classifier pays different attention to features, which means that not entire features are equally important in model training. In order to visually display the importance of features in the content marketing dataset, this paper provides comparisons among linear SVM, decision tree, and random forest. It can be seen from Figure 7 that top 10 important features that play a crucial role in model training are visually displayed. For the linear SVM, among 10 features, the first five features including pdays, poutcome_success, poutcome_failure, poutcome_nonexistent, and duration have higher impacts on linear SVM’s training process. In terms of decision tree, it is more obvious that only duration and nr.employed are the key features while other features such as euribor3m, cons.conf.idx, poutcome_success, and so on have lower importance. For the
random forest method, the top 10 features seem equal to play an important role in the training model. The relatively lower weights also indicate that most of the features are highly used.

In this paper, 20% of total samples are selected as training samples and the rest of 80% samples are treated as testing samples. Table 1 tabulates overall experimental results obtained from six classification models based on four validation metrics. As can be found from Table 1, the decision tree can capture the best predicting performances as it has the highest metric values. Compared to the decision tree, other comparison algorithms such as neural network and random forest, respectively, have the optimal results with regard to F1 score and recall. In addition, the nearest neighbor has the second-best results for F1 score, linear SVM can achieve the second-best results for recall and F1 score, the neural network has the suboptimal results for accuracy, Naive Bayes has the suboptimal results for precision, and random forest has the suboptimal results for F1 score. Execution time obtained from entire algorithms is displayed in Table 1. As can be found from this table, nearest neighbors and linear SVM require the highest computational time, while others cost much less. Figure 8 visually displays the decision boundaries of six classification methods. From those figures, we find that compared to Figures 8(a)–8(d), Figures 8(e) and 8(f) show more distinct boundaries. Figure 9 further adds the ROC curves obtained from four representative classification methods. As can be seen from Figure 9, even though the decision tree has lower area under the ROC curve than that of the nearest neighbors, it shows high intensity.

To verify the impact of different ratios of training samples on the experimental results of six algorithms, this paper performs an experiment on the content marketing

![Figure 5: Accuracy tendency in different number of depths. (a) The method using gain ratio index. (b) The method using Gini index.](image)

![Figure 6: Visualization of nodes of the decision tree method in max depth 3.](image)
**Table 1:** Comparison results obtained from six classification methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Nearest neighbor</th>
<th>Linear SVM</th>
<th>Neural network</th>
<th>Naive Bayes</th>
<th>Random forest</th>
<th>Decision tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>Recall</td>
<td>0.97</td>
<td>0.99</td>
<td>0.98</td>
<td>0.76</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.94</td>
<td>0.94</td>
<td><strong>0.95</strong></td>
<td>0.87</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.8877</td>
<td>0.8981</td>
<td>0.9037</td>
<td>0.782</td>
<td>0.8879</td>
<td><strong>0.9135</strong></td>
</tr>
<tr>
<td>Time (s)</td>
<td>3.63</td>
<td>3.07</td>
<td>0.48</td>
<td>0.07</td>
<td>0.11</td>
<td><strong>0.10</strong></td>
</tr>
</tbody>
</table>

**Figure 7:** Visualization of the importance of features in three methods.

**Figure 8:** Visualization of decision boundaries of six methods. The red and green backgrounds, respectively, denote the decision boundaries of duration and employed attribute. The red and blue solid points denote two classes corresponding to the duration and employed attributes. (a) Nearest neighbors. (b) Linear SVM. (c) Neural net. (d) Naive Bayes. (e) Random forest. (f) Decision tree.
dataset with varying the number of training samples from 5% to 25. Figure 10 provides a visual comparison of six classifiers on the different ratios of training samples. Compared to other methods, the decision tree shows the best results in all situations of training sample number. The results obtained from Naive Bayes seem disappointing primarily because it assumes that the attributes of samples are independent. In this regard, if the sample attributes are related, the effect is not good. Besides, it cannot learn the interaction among features, which highly limits its experimental performances.

4. Conclusions

Traditional content marketing strategies rely heavily on empirical knowledge such as market requirements and expectations. However, when the content marketing process meets loads of diverse users in different societies or communities, the user data will be huge so that the traditional...
content marketing strategy will be difficult. Machine learning-based methods have the ability to analyze historical data and predict future potential behaviors and activities more effectively. Among plenty of machine learning-based methods, decision tree has received more attention from content marketing managers as it intrinsically follows the process of human decision making. To verify the performance of the decision tree on the content marketing data, this paper considers a well-known decision method, called C4.5, and simultaneously compares C4.5 with other five machine learning methods including nearest neighbor, linear SVM, neural network, Naive Bayes, and random forest. Based on four validation metrics, experiments were conducted on the bank content marketing dataset under different experimental scenarios and settings. Experimental results obtained from six methods indicate that the decision tree has the ability to handle the content marketing dataset, meaning that it can provide reasonable and accurate content marketing suggestions for managers.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


