Development of Users' Call Profiles using Unsupervised Random Forest

Fatimah Almah Saaid, Robert King, Darfiana Nur
The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA

Abstract

The aim of this paper is to detect fraud in telecommunications data which consists of millions of call records generated each day. The fraud detection is implemented via the construction of user call profiles using the call detail records (CDR) data. This paper attempts to investigate the reliability of the unsupervised Random Forest method in building the profiles using its variable importance measure. Four different simulation scenarios, using different number of variable selection in each node of the tree, are performed.

Key words: Telecommunications fraud, users call profiles, unsupervised Random Forest, variable importance

1. Introduction

Global telecommunications fraud losses are estimated in the tens of billions of dollars every year [1]. Criminal acts over the telecommunications network are increasing annually which incurs millions of dollars in uncollectible debts including those due to fraud [2]. [3] reported that international fraud loss is over 20% while in the United States it is between 4% and 6%. However, the true losses are expected to be even higher since telecommunication companies are reluctant to admit fraud in their systems [4]. [5] reported that several international organizations have estimated that fraud may affect from 3% to 6% of an operator's gross revenue.

Fraud in telecommunications simply means using deceptive ways to get the network provider's service without having any intention of paying for it. There is no easy shortcut to detect fraud for many reasons. Firstly, it is due to the variation of call patterns. Subscribers' call patterns are likely to change based on factors such as promotion times, festive seasons or holidays. Secondly, fraudsters' deceptive ways vary from time to time. Previous research indicates that fraudsters are continuously attempting to change their techniques to beat the detection methods developed by service providers. As noted by [6], as soon as they realize that they are in danger of being detected, fraudsters try to invent ways to circumvent the security measures by creating new fraud techniques. Fraudsters grow wise to these countermeasures and invent new attacks such that static models of fraud lose their effectiveness [7].

According to [8], a search for fraudulent calls requires more than mere novelty of a statistical model, but it also needs fast and efficient algorithms where data mining techniques are relevant.

Data mining is often considered as a blend of statistics, artificial intelligence and data base research, which until very recently was not commonly recognized as a field of interest for statisticians [9]. There are three types of data available in telecommunications: call details records (CDR) data, network data and customer data. CDR is the most appropriate data to be considered in fraud detection. The most basic fields in CDR are the origin of number, the destination, time stamp (the date, start and end time) and call duration. While all of these are raw data, they need to be summarised into user level activities so that data mining techniques can be easily implemented. User level activities are the ones that we are interested in where the pattern of calls behaviour for normal callers can be distinguished from the fraudsters. The profiles are useful in such a way that a user's past behaviour can be captured and kept so that expected behaviour values can be developed [10]. They can be constructed by having a single numerical summary (for example maximum number of calls per week, average of day (or night) calls per week or per month) or multiple summaries (for example, mean and standard deviation of duration of call per week or per month). By constructing the profiles, a user's past behaviour can be compared with their current or future behaviour. If any inconsistency is found, this will then be
marked as indicative of fraud. Prior to achieving this aim, the selection of profiles had to be done with care because user profiles may influence the discrimination between normal and fraudulent activities [10]. To incorporate the selection, this paper proposes the use of a Random forest (RF) method to develop user call profiles for detecting telecommunications fraud. The candidate profiles will be selected using unsupervised RF’s “variable importance” measure.

The remainder of this paper is organized as follows. Section 2 consists of some brief discussions on the RF method, the unsupervised RF method and its variable importance technique. Section 3 discusses a simulation study by introducing different types of input variables such that the unsupervised RF learner may distinguish the simulated data into two distinct clusters i.e. normal and fraud activities, respectively. The simulation study uses the R packages [11] by using randomForest package [12] and MASS package [13]. This paper will be summarised with a conclusion in Section 4.

2. Random Forest

Random Forest (RF) is a machine learning algorithm that grows an ensemble of classification trees, developed by Leo Breiman (also known as the father of CART) [14]. Since its inception, the RF method has received increased attention in various applications. We choose the RF method to produce a learner for our classification problem for many reasons. Firstly, the CDR may contain more than one million call records per day and the RF is a fast-training method over large volumes of data. Secondly, the fraud events in CDR can be categorized as rare events (or unbalanced data sets) which can also be handled easily by this technique. Thirdly, the call activities need to be summarised into summary that may or may not be representative of users’ calling behaviour. The RF method can be used to produce candidate variables that are important in forming groups for normal and fraud activities. This can be achieved based on the RF predictors.

RF is a classifier consisting of a collection of $K$ tree-structured classifiers of $T_k$, $k = 1, 2, ... , K$. Each tree casts a unit vote for the most popular class at input $x$. Each $k$-th tree in the forest is constructed using different bootstrap samples with replacement. However about one-third (37%) of left out cases of the bootstrap samples are reserved and not used for constructing the $k$-th tree [15]. These samples are called out-of-bag (OOB). OOB samples are used to measure the performance of each tree where it is a form of cross-validation in predicting estimation error. This is why RF is self-testing such that pruning or cross-validation is not required.

2.1. Variable Importance

Another use of OOB samples is to estimate the variable importance in each tree. It is a measure of important predictors that are used to grow the tree. The RF method automates the selection of important variables during the tree-growing process in the forest, using the bootstrap samples. From the perspective of CDR data, determining the important profiles that can discriminate normal from fraudulent activities can be considered time consuming. We need an approach that can automate the selection of these profiles. From the RF algorithm, ideally only important variables are selected that eventually grow the tree in the forest, resulting in a learner developed with the best candidate variables.

The following steps are implemented in the RF algorithm to calculate a variable importance measure using OOB samples:

1. For each tree, $T_k$, $k = 1, 2, ... , K$ ($K$ = number of trees) with $L_k$ bootstrap samples, identify the corresponding OOB samples, $L_{k, oob}$ (OOB samples are not classified by hand but RF algorithm will select about 37% of them from the bootstrap samples).
2. Use $T_k$ to get the predicted classification model for $L_{k, oob}$.
3. Count the total number of times the tree predicts the right class (this total is referred to as a tree’s vote).
4. For each predictor $x_i$, $i = 1, 2, ... , m$, do the following:

- permute the values of $x_i$ in $L_{k, oob}$.
- use $T_k$ to get the predicted classification for $L_{k, oob}$ by using the permuted values,
- count the total number of times the tree predicts the right class, where for this study, the right class is the class label for the observed data (refer to Section 2.2),
- use the sum of votes in step 3 and deduct the sum of votes in the permuted OOB samples.

5. Averaging the difference over $K$ trees will produce the variable importance measure for each predictor.

Variable importance is one of the key innovations of the RF algorithm since it automates the selection of important candidate predictors of the learner.
2.2. Unsupervised Random Forest

CDR data is an example of unlabeled data. From a data mining perspective, it requires the unsupervised learning method to organize or cluster the CDR data into distinct groups. Fraud detection is implemented by characterizing users’ calling behaviour into two distinct clusters, normal and fraud, respectively. In data mining, unsupervised learning problems can be treated as supervised learning by using the following ideas [15]: (i) to create a class label for the observed data (the original unlabeled data) and (ii) to create another class label for synthetic data generated from a reference density. Using the combined “labeled” data, the supervised method can distinguish the observed from the synthetic, resulting dissimilarity measures. These measures can then be used as input in the unsupervised method which is implemented by using RF predictors to distinguish the observed from the synthetic data. RF predictors are an ensemble of individual tree predictors. The construction of RF trees for the combined “labeled” data yields proximity measures of each case. These measures will become the input for a clustering method such as multidimensional scaling.

For the unsupervised problem, the R package creates the synthetic data by sampling them from the product of the marginal distributions of the observed data. Denote the unlabeled data by \( x_1, x_2, \ldots, x_N \) with \( M \) variables to create a class 1 vector of \( \{x(N)\} \) observed data. To create a synthetic data set of \( N \) vectors with \( M \) dimensions (labeled as class 2), draw a sample from the one-dimensional marginal distribution of the observed data. This can be done by drawing a random sample of \( N \) values of data from each of \( M \) coordinate, yielding a sample of \( M \)-coordinate synthetic data set.

2.3. Random Forest dissimilarity measure

The dissimilarity measure is obtained from the node proximities. The RF algorithm provides proximities of two cases (data) to indicate their similarities. Suppose we have two cases \( x_i \) and \( x_j \) for \( i, j = 1, \ldots, N \) cases. If both of the cases occupy the same terminal node, their proximity, \( \text{proximity}(x_i, x_j) \), is increased by one. In the RF algorithm, the self-distance \( \text{proximity}(x_i, x_i) \) is equal to one. These proximities are calculated for all cases in the training data resulting a symmetric proximity matrix. The RF dissimilarity measure between cases \( i \) and \( j \) is given by:

\[
D(i, j) = \sqrt{1 - \text{proximity}(x_i, x_j)}. \tag{1}
\]

Multidimensional scaling is an example of the unsupervised learning method which can be done by using \( D(i, j) \) in Equation (1) as the input. In the \( R \) package this method is implemented by the \texttt{isoMDS} function [16]. To characterize the calling behaviour of normal and fraud calls from the CDR, the unsupervised RF learner shows potential in discriminating between non-fraudulent and fraudulent activities respectively, as discussed in the following section.

3. Simulations and Results

This simulation uses the \texttt{randomForest} function in \texttt{randomForest} package [12] in the \( R \) package. It is undertaken by using Equation (1) as an input to isotonic multidimensional scaling via the \texttt{isoMDS} function in \texttt{MASS} package [13]. Apart from this, the \texttt{varImpPlot} function is also implemented in searching the important input variables that are based on Gini index measure. 500 observed data are generated from mixture exponential distribution with mean duration of call of 70 seconds for normal calls and 200 seconds for fraud calls respectively. The current \( R \) 2.7.1. version uses the synthetic data sampled from the product of marginal distributions of the input variables. This simulation is designed to explore the reliability of the unsupervised RF method by using 11 variables and a different number of variables (\texttt{mtry}) to split the nodes. This simulation design is proposed to observe how sensitive the RF algorithm is in its variable selection. Four different \texttt{mtry} are used in this simulation study. Table 1 shows the run time for \texttt{randomForest} to generate 500 trees of each scenario. From Table 1, the \texttt{randomForest} function took about 5 seconds to complete. The larger the number of variables selected in each node, the longer it takes for \texttt{randomForest} function to accomplish its task.

Figure 1 displays the results of the first scenario. Although this simulation study shows there is no distinct cluster in representing the profiles for normal and fraudulent activities (Figure 1), the selection of important variables (in all four scenarios) indicates that variable 3 is the most important variable. Despite different importance ranks, the other top four candidate variables selected are 1, 2, 4 and 5. These variables are permutated from variable 1 while the rest (variable 6 - 10) are categorical variables.

4. Conclusions

The unsupervised RF method has been used to build user call profiles. This simulation study showed that, using different number of input variables used in splitting the trees’ nodes, there is no bias in the selection
process. Furthermore, identification of the first five important variables were similar regardless of the number of variables used to split the nodes.

While this is only a preliminary study, variable importance measures in the RF method were found to be reliable in rare event or unbalanced data sets, which is one of the characteristics of fraud cases. The unsupervised RF method in the simulation study indicated that the run time depends on the number of variables (mtry) used to split the nodes of each tree. Another issue arises from this study is that the multi-dimensional scaling technique did not lead to the two distinct clusters of normal and fraudulent activities. To overcome these issues, further research will be conducted either by reducing the number of trees or by applying the bootstrap sampling without replacement.

Table 1: Run time session (in R) for the simulation study of randomForest function based on different number of variables selected (mtry) in each node of the tree.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>mtry</th>
<th>Run time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>4.89</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>5.17</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5.28</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>5.49</td>
</tr>
</tbody>
</table>

References