IMPROVING TELEMARKETING INTELLIGENCE THROUGH SIGNIFICANT PROPORTION OF TARGET INSTANCES

Ding-Wen Tan  
*Universiti Tunku Abdul Rahman*, tandw@utar.edu.my

Soung-Yue Liew  
*Universiti Tunku Abdul Rahman*, syliew@utar.edu.my

William Yeoh  
*Deakin University, Australia*, william.yeoh@deakin.edu.au

Follow this and additional works at: [http://aisel.aisnet.org/pacis2014](http://aisel.aisnet.org/pacis2014)

Recommended Citation

[http://aisel.aisnet.org/pacis2014/368](http://aisel.aisnet.org/pacis2014/368)

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2014 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Abstract

In this paper we propose, develop, and test a new single-feature evaluator called Significant Proportion of Target Instances (SPTI) to handle the direct-marketing data with the class imbalance problem. The SPTI feature evaluator demonstrates its stability and outstanding performance through empirical experiments in which the real-world customer data of an e-recruitment firm are used. This research demonstrates that the feature selection using SPTI successfully improves the classifier’s performance in terms of two practical performance metrics. Additionally, we show that it outperforms other well-known feature selection methods and state-of-the-art remedies to the class-imbalance problem. Practically, the findings, when used with the classification model, will help telemarketers to better understand their customers.

Keywords: Telemarketing, Data mining, Feature selection, Binary classification, Imbalanced data.
1 BACKGROUND AND RESEARCH PROBLEM

This research was conducted in collaboration with a large online recruitment company, JobAds. One of its main businesses is selling employment posting packages which give customers credits for posting job recruitment advertisements on its website. The telesales team of JobAds requires a more accurate system for identifying and predicting customers who are likely to buy JobAds’ product; that is, identifying likely buyers. The customer-targeting system currently used by JobAds is inadequate for identifying potential customers amongst small and medium-sized enterprises (SMEs). Unlike customers of large enterprise, SMEs have relatively few employees and do not regularly hire new staff. Inevitably, this makes their purchasing pattern irregular and it is difficult to know when they may require a job-posting package. Consequently, JobAds experiences low rates of retention among such customers. This issue poses a challenge to JobAds because SMEs constitute a fairly large share of its job-posting business. In order to retain more clients among this business sector, a simple solution is to markedly increase the number of calls to the customers. But this approach has at least two drawbacks. First, more resources are needed and the operating cost will be higher if more telesales calls are made. Second, it may create a bad image of JobAds if its customers receive too many calls, especially when they do not need any job recruitment advertisement. Thus, the aim of this project is to assist JobAds’ telesales team improve its success rate by using a data-mining approach.

In regard to the problem of customer-targeting by JobAds, class imbalance is one of the main issues. This is because each month the ratio of actual buyers to non-buyers of job advertising service packages is about 1:9. The class imbalance occurs in other areas too, such as the failure of turbines in a thermal power plant (Chen et al. 2011). In this latter case failures form three different categories; normal, low, and abnormal with proportions of 67.75%, 24.65%, and 7.6% respectively. Although failure rates in the low and abnormal categories are the minority, nevertheless they are of great concern. An even more extreme case was experienced by a Canadian bank; in that instance the data set included data of a loan-product promotion where only 1.2% of clients responded from a 90,000 customers (Ling & Li 1998). Indeed, class imbalance occurs in many domains and many sectors of the economy. For example, in text classification (Makrehchi & Kamel 2007, Zheng et al. 2004), bioinformatics (Guyon et al. 2002), surveillance of nosocomial infection (Cohen et al. 2006), credit card fraud detection (Bhattacharyya et al. 2011, Duman & Ozcelik 2011), and image classification (Bhowan et al. 2009). The vital characteristic of most of the class imbalance problems is that instances in the minority class are much more valuable than the others. This is a crucial factor to be considered when applying data mining methods in order to fulfill a classification task.

The major problem caused by unbalanced class distribution is that most classifiers are generated by minimising the overall misclassification rate. Therefore, researchers may achieve good overall accuracy but the minority class may not be adequately identified. For example, referring again to the data set of the Canadian bank, if a classifier classifies all of the 90,000 customers as non-responders then it can have a very high accuracy of 98.8%, but none of the actual responders can be identified. Most prior research has led to proposals for more appropriate measures to evaluate classification outcome in the case of imbalanced data, but in those instances the research focused on the modelling, such as the modified version of existing classifiers and designs for new learning algorithms (Guo et al. 2008). In addition to these remedies, re-sampling techniques have also been proposed (Guo et al. 2008) but their effectiveness in solving class imbalance was doubted by Wasikowski and Xue-wen (2010). However, modelling is only one of the steps in a complete data mining process or in a classification task; indeed, data pre-processing has been considered to be as important (or even more important) than modelling. As indicated by Pyle (1999), Pyle (2003), Piramuthu (2004), and van der Putten (1999), data pre-processing occupies at least 80% of the time spent for completing a data mining application. In general, data pre-processing includes the tasks of dealing with incomplete, noisy, or missing values, selecting features, and transforming data. Furthermore, Pyle (1999 and 2003) claimed

---

1 For the purpose of complying with the agreement with the company, we are not able to disclose the company’s real name, and “JobAds” is used to refer to that company.
that the pre-processing stage determined 75% to 90% of the success of a data mining project, and failure of a project could likely be caused by a poor or non-existent prerequisite stage. In short, data pre-processing is a very important and necessary step in data mining because it ensures that hidden patterns of interest can be easily discovered or derived by a learning algorithm applied during modelling. Thus, the impact of data pre-processing on the class imbalance problem is the central subject of this research. Although classifications for direct marketing always suffer from a class imbalance, limited research has been done for tackling the imbalance experienced by direct marketing businesses. Thus, this research focuses on feature-selection, a factor which usually receives less attention than other data pre-processing approaches such as re-sampling. Although feature selection is a relatively new remedy for addressing class imbalance, its classification outcomes have been noticeably improved by using feature selection, as reported by Huanjing et al. (2010), Wasikowski and Chen (2010), and Zhang et al. (2012).

In this study we propose and develop a new single-feature evaluator called Significant Proportion of Target Instances (SPTI) which is designed to use direct-marketing data which contains a class imbalance problem. For each feature the SPTI can measure the proportion of minority-class instances (target instances) in each feature value. The justification for this project is that minority-class instances are a particular concern for direct marketers. This paper is organised as follows. In section 2, we compare and discuss various feature selection methods and introduce the single feature evaluator that could help classifier achieve better performance especially in the presence of a class imbalance. In section 3, we describe the experiment framework and provide details about the experiments. The complete results of the experiments are provided in section 4. Finally, section 5 concludes the study and suggests directions for future research.

2 FEATURE SELECTION METHOD FOR THIS RESEARCH

In this section, certain effective and well-known feature selection methods are investigated. Also, we propose and explain a new feature selection method (SPTI) which is developed particularly for data with a class imbalance.

2.1 Feature Subset Selection Methods

Two well-known feature subset selection methods used in data mining have been applied in this study; they are correlation-based feature selection (CFS) and subset consistency (SC). According to Hall (2000), CFS measures the worth of a set of features by assessing the predictive ability of every feature separately while taking into account the redundancy level among them. CFS is always searching for an ideal feature subset that comprises features that are highly correlated with the target feature but which are not intercorrelated (or slightly intercorrelated). In addition, Hall (2000) has shown that CFS could significantly reduce the number of input features while preserving or improving the performance of a predictive model.

On the other hand, SC measures the degree of inconsistency of a feature subset with respect to the target values of the target feature. According to Liu and Setiono (1996), if two instances hold exactly the same values for all features but have different target values, they are considered to be inconsistent. For each group of inconsistent instances, the inconsistency count is calculated by subtracting the largest number of instances that hold the same target values from the number of the instances. The inconsistency rate of a feature subset is the sum of all the inconsistency counts divided by the total number of instances. Since the full feature set always has the highest consistency, SC can be used to locate the smallest feature subset which has the same consistency as that of the full feature set. Liu and Setiono (1995) and Liu and Setiono (1996) have reported that SC was able to reduce the number of features significantly and they have also shown that the performances of two learning algorithms (i.e. ID3 and C4.5) for generating decision trees were improved after using SC.

Our data set consists of 37 features so it is impractical to use an exhaustive search because there are $2^{37}$ combinations of feature subsets. Since an exhaustive search is computationally prohibitive, a heuristic search is an alternative. In this study, backward elimination is used as a searching method for
both feature subset evaluators because backward elimination is a better choice for feature subset selection than forward selection (Guyon & Elisseeff, 2003). Forward selection tends to find weaker feature subsets because the value of selected subsets is not assessed in the presence of other features that have not yet been included. Nilsson et al. (2007) also noted that backward feature elimination is better than forward selection in terms of consistency. Indeed, backward elimination for feature selection has been successfully applied in the studies conducted by Pellet and Elisseeff (2008), Cho (2009), Feng et al. (2011), and Inbarani and Banu (2012).

2.2 Significant Proportion of Target Instances (SPTI)

Many techniques have been proposed and used for feature selection. However, there has been limited research on feature selection for binary classification of imbalanced data. Binary classification is commonly applied and is thus an important data mining issue. For example, classifying patients and non-patients for medical studies (Yu et al. 2012), and classifying churners and non-churners of a service for a telecommunication company (Huang et al. 2012).

Chawla et al. (2004) observed that most real-world class imbalances require ad hoc solutions because different data mining tasks consist of different data types and requirements. Also, it is understandable that different data mining problems need different feature selection methods as indicated by Dash and Liu (1997) and Liu and Yu (2005). It is therefore necessary to develop a feature selection method suitable for binary classification of imbalanced data. This particularly applicable to the class imbalance problem and a novel single feature evaluator (or feature ranking method) based on preference for target instances is proposed and presented in this section. Single feature evaluator is a type of filter model, its major advantages being fast computation, flexibility, and practicality - which have been shown in previous research and real-world applications (Yuan et al. 2009; Lihong et al. 2011; Beigy & Sadeghi 2013).

In particular, the idea of SPTI is based on the following reasoning. First, in a binary classification problem with class imbalance, correct classification of a positive instance (target instance) is much more important than correct classification of a negative instance but the proportion of positive instances is significantly less than that of negative instances. Therefore, it is more difficult to correctly recognise positive instances, yet positive instances should be given more attention. Second, if the feature values of positive instances vary considerably and the distribution of positive instances is relatively scattered in the presence of negative instances for a certain feature, it brings difficulty to learn the pattern of positive instances. Thus, removing such unwanted feature is necessary. SPTI is designed to evaluate the significance degree of the proportion of positive instances, and rank features according to SPTI measure so that unwanted features can be identified.

2.3 Algorithm of Filter-Based Feature Selection Model Using SPTI

Given a training data set, the algorithm of our filter-based feature selection model using SPTI consists of four steps. First, it counts $N_f^i$ and $N_{f1}^i$, which are the total number of instances and the number of target instances in value $i$ of feature $f$, respectively. Second, it computes SPTI for every feature. Third, it weights and ranks the features according to their SPTIs. Fourth, it selects the desirable features from the ranked list based on a predetermined threshold. SPTI is shown by F1.

$$SPTI(f) = \sum_{i=1}^{R_f} \frac{N_{f1}^i}{N_f^i} \left[ (N_{f1}^i - E_{f1}^i) + \left| N_{f1}^i - E_{f1}^i \right| \right]$$  \hspace{1cm} (F1)

where $R_f$ is the number of feature values of feature $f$, and $E_{f1}^i$ is the expected number of target instances in value $i$ of feature $f$.

Particularly, in the fourth step, a threshold $h$ is used to select the desirable features from the full feature set. Assume there is a benchmark feature $f_b$ possessing the following properties.
There are only two feature values and they separate all the instances equally, that is, \(N_{1}^{f_{b}} = N_{2}^{f_{b}} = N/2\), where \(N\) is the total number of instances.

The proportion of target instances with the first feature value \(\frac{N_{11}^{f_{b}}}{N_{1}^{f_{b}}}\) is \(h\) times of the ratio of, \(\frac{N_{pos}}{N}\), that is, \(\frac{N_{11}^{f_{b}}}{N_{1}^{f_{b}}} = h \frac{N_{pos}}{N}\), where \(N_{pos}\) is the total number of target instances, and \(h\) is greater than 1.

The second feature value of \(f_{b}\) will actually be ignored in calculating SPTI\((f_{b})\) since the second property makes \(N^{f_{b}}_{2} < E^{f_{b}}_{2}\). Thus, SPTI\((f_{b})\) is simply equal to \(h(h - 1)(N_{pos})^2/N\). Any feature whose SPTI is less than SPTI\((f_{b})\) will be excluded and the remaining features will form the selected feature subset.

One way to determine an appropriate value for \(h\) is to use Pearson's chi-squared test. By varying the significant level \(\alpha\) for Pearson's chi-squared test, a different \(f_{b}\) with a different confidence level on the dependency with target feature can be defined. For example, if \(\alpha = 0.05\), there is 95% confidence that the resulting \(f_{b}\) is significantly related to the corresponding target feature. Note that the degree of freedom is equal to 1 because of the setting of \(f_{b}\). In the experiments conducted for this study, three different values 0.005, 0.05, and 0.5 have been assigned to \(\alpha\) in order to investigate the impact of \(\alpha\) on different classifiers. The \(\alpha\) values 0.005, 0.05, and 0.5 approximately correspond to \(h\) values 1.0298, 1.0208, and 1.0072, respectively. Based on a given \(\alpha\) value, assume that the corresponding chi-squared value is \(C_{\alpha}\) for one degree of freedom, then the \(h\) value can be calculated by F2.

\[
h = 1 + \sqrt{\frac{C_{\alpha}(N - N_{pos})}{N^2 N_{pos}}} \quad (F2)
\]

3 EXPERIMENT SET-UP

This section provides details on how the experiments were designed for this study.

3.1 Data set

A real-world data set with a class imbalance derived from the direct-marketing sector was used in study. The data were provided by a large online employment advertising company – JobAds, which is one of the largest online recruitment companies in Southeast Asia. A typical service provided by JobAds is online job-advertisement posting on its web sites. However, JobAds has suffered from low rates of retention of its SME customers. In the past, the telesales team at JobAds would call only 10 percent of all SME customers due to resource constraints, such as the limited manpower and budget for calling expenses. Rather than cold-calling individual SME customers, currently classifiers are generated to identify likely buyers of the product. This can help the telesales team in JobAds to focus on those potential SME customers.

The data set currently describing JobAds clients contains 37 features and one response variable representing class label. The response variable indicates if a customer purchased at least one product in a certain month. That variable has two classes: one is “Yes” which means a customer buys at least one product, and the other is “No” which means a customer does not buy any product. “Yes” customers are target instances (or positive instances) while “No” customers are negative instances. A ten-fold cross-validation was applied to the data set.

3.2 Experiment Objective and Framework

The first experiment tested whether the feature selection using SPTI could compete with two well-known feature selection methods, correlation-based feature selection (CFS) and subset consistency (SC) feature selection. The second experiment examined the effectiveness of the re-sampling technique and the ensemble learning strategy in addition to feature selection. It investigated which remedy to the class imbalance is most effective. Specifically, random under-sampling and rotation-
based ensemble learning were selected due to their effectiveness and outstanding performance as shown by Raskutti and Kowalczyk (2004), Al-Shahib et al. (2005), Van Hulse et al. (2007), De Bock and Poel (2011), and Budnik et al. (2012). In particular, by means of extensive study and by empirical experiments Van Hulse et al. (2007) showed that random under-sampling is the best re-sampling approach.

3.3 Learning Algorithms

The main objective of this study has been to investigate the ability of feature selection in reducing the negative influences associated with the class imbalance problem. Comparison between the performance of a classifier without any feature selection and that of a classifier using certain feature-selection technique was necessary for it is well known that classifiers generated by different learning algorithms have different performances even though they have the same training set; that is, they have the same instances and features. Therefore, we tested the selected feature set on different types of classifiers with different induction mechanisms.

Artificial neural networks (ANNs), logistic regression (LR), and support vector machines (SVMs) are used to build classifiers, and they have been used widely in machine learning and data mining. Two open-source tools, KNIME (2.5.1) (Berthold et al. 2007) and Waikato Environment for Knowledge Analysis (WEKA) (3.6) (Hall et al. 2009), were deployed to complete all experiments. We also used KNIME to perform some data cleaning, and we calculated the area under the lift chart curve (AULIFT). We also used WEKA to run the above-mentioned feature evaluators and the learning algorithms of ANN and LR. To generate the SVM classifier, the sophisticated software SVM-perf, developed by Joachims (2005), was used in this study.

3.4 Performance Evaluation

The two evaluation measures adopted in the experiments are represented by the area under the lift chart curve (AULIFT) and the number of “Yes” customers (actual buyers) out of the 1000 customers identified by the classifier (NYC_{1000}). The AULIFT assesses the overall performance of the classifier and the NYC_{1000} assesses the classifier in a practical manner. Usually, the telesales team at JobAds would target and call only a few thousands of all the SME customers per month due to resource constraints.

Witten et al. (2011) stated that a lift chart curve is plotted by the number of true responses versus the number of customers contacted. In the data mining research for direct marketing, lift chart is always used to provide a clear picture of the real effectiveness of a classifier and how to determine the appropriate number of customers to contact in order to achieve the target number of true responses. Moreover, we can readily evaluate and compare the overall performance of different classifiers by AULIFT. In brief, the rule for comparison is that the bigger the AULIFT the better the performance.

Recent research on feature selection has also considered stability as an important element, researchers being keen to develop stable feature selection methods (Kalouhstis et al. 2007; He & Yu 2010; Alelyani et al. 2011). Feature selection stability is used to analyse whether a feature selection method will choose the same feature subset or whether it will give a feature subset with little change when perturbation occurs in the training set. Several metrics have been proposed to measure and quantify the stability of a given feature selection method. In general, the metrics can be categorised into three groups. In the first group they are used to evaluate feature selection methods that can assign a weight or score to each feature according to its importance. For example, Pearson’s correlation coefficient (Kalouhstis et al. 2007). In the second group the metrics can be used to measure the similarity of the feature-rank lists generated by the feature selection method. Examples are Spearman’s rank correlation coefficient (Kalouhstis et al. 2007) and Canberra distance (Jurman et al. 2008). In the third group the metrics can be used to evaluate the similarity of the feature subsets selected by the feature selection method. For example, Dice-Sorensen’s index (Yu et al. 2008: He & Yu 2010) and the Jaccard index (Kalouhstis et al. 2007; Alelyani et al. 2011). Kalouhstis et al. (2005) and Kalouhstis et al. (2007) reported that the Jaccard index gave the most important information about the feature selection stability in comparison with Pearson’s correlation coefficient and Spearman’s rank correlation.
coefficient. In their experiment, Pearson’s correlation coefficient and Spearman’s rank correlation coefficient were likely to overestimate the feature selection stability. Thus, the Jaccard index was selected for this study. Equation E1 provides the formula for calculating the Jaccard index for two selected feature subsets. Equation E2 shows the extension of the Jaccard index for measuring the selection stability for more than two selected feature subsets; for example, 10-fold cross validation.

The value of the Jaccard index is in the interval [0, 1] where 1 means all the considered feature subsets are identical, and 0 means no common features at all.

\[
\text{Jaccard index} = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}
\]

(E1)

\[
S_{\text{Jaccard}} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{|A_i \cap A_j|}{|A_i \cup A_j|}
\]

(E2)

Therefore, in addition to the AULIFT and the NYC\(_{1000}\) the stability of each feature selection method on their selected feature subset for the 10-fold cross validation will also be examined using the Jaccard index.

4 RESULT

In this work the data set used includes the purchasing records within the second half of the year 201x (the year has been disguised because real-world data has been used in this research); that is, Ord-Jul_to_Dec-201x. And 10-fold stratified cross validation is applied to test the effectiveness of the feature selection methods.

4.1 Comparison of Different Feature Selection Methods

Table 1 shows the average number of features selected by different feature subset evaluators for the 10-fold cross validation. Note that SPTI-0.005, SPTI-0.05 and SPTI-0.5 mean that significance levels are set to be 0.005, 0.05 and 0.5, respectively. NoF refers to the number of features, and the full feature set is represented by FullFS. Table 2 provides the average AULIFT and the average NYC\(_{1000}\) achieved by the classifiers with different feature subset evaluators when using ANN, LR, and SVM respectively.

In this comparison, SPTI-0.005, SPTI-0.05, and SPTI-0.5 give the best performance for ANN, logistic regression, and SVM respectively. It shows that, with an appropriate significant level or threshold \(h\), SPTI can select a good feature subset that improves the performance of a classifier in terms of both AULIFT and NYC\(_{1000}\). In addition, SPTI surpasses CFS and SC based on the average performances over this 10-fold cross-validation.

In this experiment, the ANN classifier performs poorly when the selected feature subset is relatively large. For example, the feature subsets selected by SPTI-0.05, SPTI-0.5, FullFS and SC. This implies that the ANN classifier is more sensitive to irrelevant features compared to LR and SVM classifiers.

Table 2 shows that both SPTI-0.005 and CFS can improve the performance of the ANN classifier; moreover SPTI-0.005 helps the ANN classifier to achieve the best AULIFT.

The proportion of “Yes” customers (target instances) of the data set is only 10.74%. That means that the expected NYC\(_{1000}\) from the use of random customer targeting is about 107.4. All the classifiers manage to produce a better NYC\(_{1000}\). In particular, the best NYC\(_{1000}\) achieved by the SVM classifier with SPTI-0.5 is 190.3; this means that the number of successfully identified “Yes” customers is increased by about 77.2% from 107.4 to 190.3. This provides strong evidence that data mining and SPTI are useful for tackling customer targeting problems in direct marketing.
<table>
<thead>
<tr>
<th>Method</th>
<th>NoF</th>
<th>SPTI-0.005</th>
<th>SPTI-0.05</th>
<th>SPTI-0.5</th>
<th>FullFS</th>
<th>CFS</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>14.3</td>
<td>20.2</td>
<td>28.1</td>
<td>37.0</td>
<td>4.0</td>
<td>35.0</td>
</tr>
</tbody>
</table>

Table 1. Average number of features selected by certain feature subset evaluator for the 10-fold cross validation.

<table>
<thead>
<tr>
<th>Method</th>
<th>AULIFT (ANN)</th>
<th>AULIFT (LR)</th>
<th>AULIFT (SVM)</th>
<th>NYC1000 (ANN)</th>
<th>NYC1000 (LR)</th>
<th>NYC1000 (SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPTI-0.005</td>
<td>0.6211</td>
<td>0.6139</td>
<td>0.6041</td>
<td>187.9</td>
<td>185.1</td>
<td>185.9</td>
</tr>
<tr>
<td>SPTI-0.05</td>
<td>0.5468</td>
<td>0.6153</td>
<td>0.6045</td>
<td>148.9</td>
<td>189.5</td>
<td>186.7</td>
</tr>
<tr>
<td>SPTI-0.5</td>
<td>0.5331</td>
<td>0.6113</td>
<td>0.6052</td>
<td>137.4</td>
<td>187.1</td>
<td>190.3</td>
</tr>
<tr>
<td>FullFS</td>
<td>0.5826</td>
<td>0.6117</td>
<td>0.6047</td>
<td>176.0</td>
<td>187.2</td>
<td>189.3</td>
</tr>
<tr>
<td>CFS</td>
<td>0.6109</td>
<td>0.6059</td>
<td>0.5985</td>
<td>184.0</td>
<td>183.3</td>
<td>182.4</td>
</tr>
<tr>
<td>SC</td>
<td>0.5831</td>
<td>0.6117</td>
<td>0.6049</td>
<td>175.2</td>
<td>187.3</td>
<td>190.0</td>
</tr>
</tbody>
</table>

Table 2. Average performances of different classifiers with different feature subset evaluators.

Next, statistical tests were conducted to check if the AULIFT and NYC1000 gained by the classifiers using the feature subset evaluators, and if the gains are statistically greater than those achieved without the use of the feature selection. More specifically, parametric and nonparametric paired-sample tests (paired samples t-test and Wilcoxon signed-rank test respectively) were used to examine if the improvement on the classifier’s performance by using feature selection is statistically significant.

Table 3 shows the results (p-values) of left-tailed paired sample t-tests on the differences between the performances of the classifiers using different feature selections for ANN, LR, and SVM classifiers respectively. At the 5% significance level (α = 0.05), in this experiment, Table 3 shows that the improvements on AULIFT and NYC1000 made by SPTI-0.005 for the ANN classifier are statistically significant. CFS was only able to significantly improve the AULIFT for the ANN classifier. The rest of the feature subset evaluators do not provide any significant improvement to the AULIFT and NYC1000. For the LR classifiers, Table 3 shows that only SPTI-0.05 was able to significantly improve the AULIFT. For NYC1000, although the p-value achieved by SPTI-0.05 is the smallest (the best), it is greater than 0.05. This means that none of the feature subset evaluators yielded significant improvements. For the SVM classifiers, Table 3 shows that SC and SPTI-0.5 gave the smallest p-value for the improvements on AULIFT and NYC1000 respectively, but the improvements were not statistically significant.

Table 4 shows the results (p-values) of the Wilcoxon signed-rank test on the difference between the performances of the classifiers using different feature selections based on positive ranks for ANN, LR, and SVM classifiers, respectively. The results of the Wilcoxon signed-rank tests tally with the findings based on the paired samples t-tests.

To summarise, based on the overall performance, SPTI is able to reduce the computational complexity by excluding irrelevant features. Not only can computational time be reduced but, more importantly, the classifier’s performance can be markedly improved, especially for the ANN and LR classifiers. In this experiment, both the existing feature subset evaluators, CFS and SC, cannot compete with SPTI because they failed to provide significant improvements for the classifiers – and in some cases their performances were worse. On the other hand, at the 5% significance level both the paired samples t-test and the Wilcoxon signed-rank test showed that there are no statistically significant differences among the highest scores of NYC1000 achieved by the ANN, logistic regression, and SVM classifiers. However, the largest AULIFT gained by the ANN classifier is statistically better.
than the ones gained by the logistic regression and SVM classifiers. That means that ANN is the learning algorithm best suited to this experiment.

### Table 3. Results (p-values) of the left-tailed paired samples t-tests for different classifiers (p-values less than 0.05 indicate significant difference and are highlighted in bold).

<table>
<thead>
<tr>
<th>Difference</th>
<th>AULIFT (ANN)</th>
<th>AULIFT (LR)</th>
<th>AULIFT (SVM)</th>
<th>NYC\textsubscript{1000} (ANN)</th>
<th>NYC\textsubscript{1000} (LR)</th>
<th>NYC\textsubscript{1000} (SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FullFS – SPTI-0.005</td>
<td>0.0000</td>
<td>0.1009</td>
<td>0.1009</td>
<td>0.0137</td>
<td>0.7943</td>
<td>0.9228</td>
</tr>
<tr>
<td>FullFS – SPTI-0.05</td>
<td>0.9990</td>
<td>0.0313</td>
<td>0.0313</td>
<td>0.9962</td>
<td>0.1596</td>
<td>0.9330</td>
</tr>
<tr>
<td>FullFS – SPTI-0.5</td>
<td>1.0000</td>
<td>0.8064</td>
<td>0.8064</td>
<td>0.9999</td>
<td>0.5514</td>
<td>0.1582</td>
</tr>
<tr>
<td>FullFS – CFS</td>
<td>0.0000</td>
<td>0.9885</td>
<td>0.9885</td>
<td>0.0556</td>
<td>0.8364</td>
<td>0.9808</td>
</tr>
<tr>
<td>FullFS – SC</td>
<td>0.4488</td>
<td>0.4771</td>
<td>0.4771</td>
<td>0.5785</td>
<td>0.4440</td>
<td>0.2980</td>
</tr>
</tbody>
</table>

### Table 4. Results (p-values) of the Wilcoxon signed-rank test for different classifiers (p-values less than 0.05 indicate significant difference and are highlighted in bold).

<table>
<thead>
<tr>
<th>Difference</th>
<th>AULIFT (ANN)</th>
<th>AULIFT (LR)</th>
<th>AULIFT (SVM)</th>
<th>NYC\textsubscript{1000} (ANN)</th>
<th>NYC\textsubscript{1000} (LR)</th>
<th>NYC\textsubscript{1000} (SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FullFS – SPTI-0.005</td>
<td>0.0010</td>
<td>0.2461</td>
<td>0.6523</td>
<td>0.0186</td>
<td>0.7891</td>
<td>0.9355</td>
</tr>
<tr>
<td>FullFS – SPTI-0.05</td>
<td>0.9971</td>
<td>0.0420</td>
<td>0.7217</td>
<td>0.9932</td>
<td>0.2617</td>
<td>0.9082</td>
</tr>
<tr>
<td>FullFS – SPTI-0.5</td>
<td>0.9990</td>
<td>0.8389</td>
<td>0.1875</td>
<td>0.9990</td>
<td>0.5059</td>
<td>0.1406</td>
</tr>
<tr>
<td>FullFS – CFS</td>
<td>0.0010</td>
<td>0.9756</td>
<td>0.9951</td>
<td>0.0586</td>
<td>0.8145</td>
<td>0.9775</td>
</tr>
<tr>
<td>FullFS – SC</td>
<td>0.4609</td>
<td>0.5000</td>
<td>0.1377</td>
<td>0.6426</td>
<td>0.4570</td>
<td>0.1992</td>
</tr>
</tbody>
</table>

Table 5 provides the Jaccard index \( S_{\text{jacc}} \) for each feature subset selection method in order to measure their selection stability. Since both CFS and SC selected the same feature subset for all the 10-fold cross validation data sets, their \( S_{\text{jacc}} \) is equal to 1. Each of the three SPTI subset selection methods had a very high score on the Jaccard index. Based on the scores, SPTI-0.005, SPTI-0.05, and SPTI-0.5 managed to select 12.8 common features out of 14.3 features, 19.6 common features out of 20.2 features, and 27.5 common features out of 28.1 features respectively. The average number of different features in the selected feature subset (denoted by \( n_{\text{diff}} \)) is also given in Table 5. Here, the largest \( n_{\text{diff}} \) produced by SPTI is only 1.5 which means that, on average, no more than two different features occurred in the feature subsets. It shows that SPTI can improve the classifier’s performance as well as provide a stable feature selection method.

<table>
<thead>
<tr>
<th>Method</th>
<th>( S_{\text{jacc}} )</th>
<th>( n_{\text{diff}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPTI-0.005</td>
<td>0.8971</td>
<td>1.5</td>
</tr>
<tr>
<td>SPTI-0.05</td>
<td>0.9683</td>
<td>0.6</td>
</tr>
<tr>
<td>SPTI-0.5</td>
<td>0.9796</td>
<td>0.6</td>
</tr>
<tr>
<td>CFS</td>
<td>1.0000</td>
<td>0.0</td>
</tr>
<tr>
<td>SC</td>
<td>1.0000</td>
<td>0.0</td>
</tr>
</tbody>
</table>

### Table 5. Jaccard index and average number of different features selected for the 10-fold cross validation.

#### 4.2 Comparison with Other Remedies

Figure 1 and 2 compare the average AULIFT and NYC\textsubscript{1000} for the 10-fold cross validation achieved by different remedies to the class imbalance problem. Seven classifiers were built with different individual remedies. They are SPTI (with \( \alpha = 0.005 \)), random under-sampling (RUS), rotation-based ensemble learning (REL), and also the hybrid approaches: SPTI feature selection paired with random under-sampling (F&S), SPTI feature selection paired with rotation-based ensemble learning (F&L), random under-sampling paired with rotation-based ensemble learning (S&L), and the combination of all three remedies (ALL). The baseline performance, which is done by the classifier without any of the three remedies, is represented by a horizontal line in both figures 1 and 2. Note that the learning...
algorithm used in this section is artificial neural networks; this is used because the previous experiment showed that the ANN classifier is the best choice for the data being applied in for this study.

The rankings of every remedy are the same for both AULIFT and NYC\textsubscript{1000}. The best performances were achieved by using SPTI alone in terms of AULIFT and NYC\textsubscript{1000}. This contradicts the finding of Kehan et al. (2012) in which hybrid approaches performed better than feature selection used alone. F&L was the second-best performer. The others gave poor results that are far worse than SPTI – some being even lower than the baseline performances. This shows that neither a training set with a balanced class distribution nor ensemble learning can help to generate an effective classifier for a class imbalance problem. Also, this supports the theory that reducing the number of negative instances may lose useful information and eventually downgrade the classifier’s performance (Manevitz & Yousef 2002; Wasikowski & Chen 2010). On the other hand, an interesting finding is that both F&S and F&L beat their counterparts RUS and REL respectively. This means that SPTI was able to improve the performance of the classifier using random under-sampling or rotation-based ensemble learning.

![Figure 1. The AULIFT achieved by different remedies.](image1)

![Figure 2. The NYC\textsubscript{1000} achieved by different remedies.](image2)

Even though the averages of both AULIFT and NYC\textsubscript{1000} which were obtained by the classifiers using SPTI are higher than the other approaches, we must examine if the differences are statistically significant. Tables 6 and 7 show the results (p-values) of the left-tailed paired samples t-tests and the Wilcoxon signed-rank test on the differences based on positive ranks. According to the parametric paired samples test shown in Table 6, all the p-values are less than 0.05 so we can conclude that the performances of SPTI, in terms of both AULIFT and NYC\textsubscript{1000}, are significantly better than all other six approaches at the 5% significant level. On the other hand, the nonparametric paired samples test shown in Table 7 tells that at the 5% significant level SPTI significantly outperformed the other six approaches in terms of both AULIFT and NYC\textsubscript{1000}; however, the difference between the NYC\textsubscript{1000} gained by SPTI and F&L is not so statistically significant. Nevertheless, the p-value of the difference
between the NYC\textsubscript{1000} gained by SPTI and F&L is 0.0508 (5.08%), which is very close to 0.05, and it means that the different is still quite large.

<table>
<thead>
<tr>
<th>Difference</th>
<th>AULIFT</th>
<th>NYC\textsubscript{1000}</th>
</tr>
</thead>
<tbody>
<tr>
<td>F&amp;L – SPTI</td>
<td>0.0003</td>
<td>0.0229</td>
</tr>
<tr>
<td>ALL – SPTI</td>
<td>0.0000</td>
<td>0.0009</td>
</tr>
<tr>
<td>REL – SPTI</td>
<td>0.0000</td>
<td>0.0052</td>
</tr>
<tr>
<td>S&amp;L – SPTI</td>
<td>0.0000</td>
<td>0.0002</td>
</tr>
<tr>
<td>F&amp;S – SPTI</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>RUS – SPTI</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 6. Results (p-values) of the left-tailed paired samples t-tests (p-values less than 0.05 indicate significant difference and are highlighted in bold).

<table>
<thead>
<tr>
<th>Difference</th>
<th>AULIFT</th>
<th>NYC\textsubscript{1000}</th>
</tr>
</thead>
<tbody>
<tr>
<td>F&amp;L – SPTI</td>
<td>0.0020</td>
<td>0.0508</td>
</tr>
<tr>
<td>ALL – SPTI</td>
<td>0.0010</td>
<td>0.0029</td>
</tr>
<tr>
<td>REL – SPTI</td>
<td>0.0010</td>
<td>0.0068</td>
</tr>
<tr>
<td>S&amp;L – SPTI</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td>F&amp;S – SPTI</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td>RUS – SPTI</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Table 7. Results (p-values) of the Wilcoxon signed-rank test (p-values less than 0.05 indicate significant difference and are highlighted in bold).

Like other studies this project has limitations. The proposed SPTI handles only discrete values, and in order to work on those features with continuous data, the data need to be pre-processed by discretisation or binning. Additionally, the information-loss caused by the discretisation may influence the classifier’s performance. It can have negative effects if useful information is lost or it may yield benefits if noise is removed.

5 CONCLUSION

This paper proposes an effective feature evaluator to facilitate feature selection for binary classification of imbalanced data. Specifically, a feature evaluator SPTI was designed to enhance the classifiers which suffer from a class imbalance. This study used real-world data sets; it demonstrated that feature selection using SPTI is able to improve a classifier’s performance in terms of AULIFT and NYC\textsubscript{1000}, and it also showed that SPTI outperforms two other well-known feature selection techniques. Moreover, the stability and robustness of the SPTI has been proven in this study.

Two key findings emerged from this research. First, a feature subset evaluator may not necessarily outperform a single feature evaluator, although it performs a more complete search in the feature space. Rather than spending computational resources on searching the best feature subset, this study shows that more attention should be paid to identifying those features which have highest discrimination power. Second, instead of manipulating the data distribution or using ensemble learning, feature selection is much more useful for tackling a class imbalance.

The practical contribution of this study is that it addressed a real-world telesales problem being faced by a large online employment advertising company by using classification with feature selection. Specifically, it has tackled the challenge of targeting SME customers within an online job advertising context. This study also demonstrates that the classifier using SPTI can serve as an effective customer-targeting tool which can assist firms like JobAds to identify potential SME customers. JobAds has benefited from improved retention rates among its SME customers through the proposed classification with feature selection. Thus, it shows that data mining and feature selection techniques
can help managers make better decisions by providing useful predictive analysis and by reducing human bias. By improving the customer-retention rate with the help of data mining and feature selection, the direct-marketing business can be markedly improved in terms of profit and market share. As for future research, the applicability of SPTI in other areas, such as text categorisation, image processing, and bioinformatics is suggested. Further research into the behaviour and performance of SPTI using other data sets would be interesting and useful.

References


