Neural Comput & Applic (2017) 28:2593–2603
DOI 10.1007/s00521-016-2214-y

ORIGINAL ARTICLE

A new transferred feature selection algorithm for customer identification

Bing Zhu[1] · Yongge Niu[1] · Jin Xiao[1] · Bart Baesens[2,3]

Received: 11 May 2015 / Accepted: 19 January 2016 / Published online: 2 February 2016

Abstract Class imbalance brings great challenges to feature selection in customer identification, and most of the current feature selection approaches cannot produce good prediction on the minority class. A number of studies have attempted to solve this issue by using resampling techniques. However, resampling techniques only use the in-domain information and they cannot achieve good performance when the imbalance is caused by the absolute rarity of the minority class. In this paper, we focus on the issue of feature selection with class imbalance caused by absolute rarity. By introducing the idea of transfer learning, we develop a transferred feature selection method based on the group method of data handling neural networks. The proposed ensemble neural network extracts information of similar customers from related domains to deal with the information scarcity of the minority class in the target domain. Experiments are done on a real-world application using data from a cigarette company. The results indicate that the new method gives better predictive performance than other benchmark feature selection methods, especially in terms of the predictive accuracy of the minority high-value customers. At the same time, the new algorithm can help to identify important features that distinguish high-value customers from low-value ones.

Keywords Customer identification · Feature selection · Class imbalance · Absolute rarity · Transfer learning · Group method of data handling

1 Introduction

Customer identification is the first stage of customer relationship management (CRM), which involves finding and targeting the customers who are most profitable to companies [29]. Feature selection is a useful tool in this process, and it will not only solve the "curse of dimensionality" problem in modeling, but also help companies to identify a key feature subset that distinguishes high-value customers from low-value ones [24]. Therefore, feature selection has received considerable attention in customer identification.

The data used for customer identification are often imbalanced. There are usually much less high-value customers (minority class) than low-value ones (majority class). The class imbalance issue brings great challenges to current feature selection methods. Most of them tend to find features that are apt to classify all customers as belonging to the majority class, which results in high overall accuracy but unacceptably low accuracy with respect to the minority class. This would be useless at detecting the minority high-value customers. Some people have attempted to solve this issue by using resampling techniques [13, 40]. However, resampling methods do not

✉ Yongge Niu
niuyongge@gmail.com

Bing Zhu
zhubing1866@hotmail.com

Jin Xiao
xiaojin@scu.edu.cn

Bart Baesens
bart.baesens@kuleuven.be

[1] Business School, Sichuan University, Chengdu 610064, People's Republic of China

[2] Department of Decision Sciences and Information Management, KU Leuven, 3000 Leuven, Belgium

[3] School of Management, University of Southampton, Highfield, Southampton SO17 1BJ, UK

add any new information to the minority class. So they are useful only if enough data are provided for both classes and could fail when the class imbalance is caused by absolute rarity of the minority instances (called absolute imbalance). Absolute imbalance refers to the situation where the minority class has a very small sample size [43]. For most CRM applications massive customer databases are available due to the development of information technology, but there are still many situations in which only a small number of data are available. For example, sometimes data collection through surveys or experiments is very expensive in terms of both time and economic cost, and firms can only obtain a small sample size of customer data. New services and products are other sources of small data size, where records of customer behavior are scarce, especially records of high-value customers. Hence, it is necessary to develop some feature selection methods that could deal with class imbalance caused by absolute rarity.

In CRM practice, abundant data of similar customers from other domains are often available. For example, similar customer data might have been gathered at an earlier time or in other areas with the same purpose. Clearly, customer data from related domains provide valuable information, especially information of the minority high-value customers. As Japkowicz and Stephenin have shown [19], relatively more information about the minority class benefits the classification modeling, which is then able to distinguish rare minority instances from the majority ones. Thus, these data from a related domain could help to alleviate the problem of class imbalance due to the shortage of the minority samples in the domain of interest (hereafter, referred to as the target domain, and the related domain is referred to as the source domain). However, data distributions in different domains are different. Thus, it is an interesting research topic to find suitable ways to utilize the information in the source domain. Recently, the so-called transfer learning technique has been proposed to exploit source domain data in a principled manner [32].

Transfer learning is an emerging family of machine learning techniques. It applies the knowledge learned in related domains to develop efficient models in the target domain [4]. It has been proposed to deal with situations where data are difficult to collect in the target domain while auxiliary data in the source domain are readily available or relatively easy to collect. Previous research has shown that transfer learning can improve modeling accuracy in many areas such as text mining [11], image analysis [27]. To the best of our knowledge, no reported research has applied transfer learning to solve the class imbalance issue in feature selection.

In this paper, we focus on the issue of feature selection with class imbalance caused by absolute rarity. We try to introduce the idea of transfer learning into feature selection and help to relieve information scarcity of the minority class by integrating the auxiliary data from the source domains. For this purpose, an ensemble of group method of data handling neural networks (GMDH) is developed. The new method, called transferred feature selection based on GMDH (TFSG), first creates multiple training datasets incorporating both source and target data via sampling and then uses GMDH to select feature subsets for each training dataset. One advantage of GMDH is that it will automatically select significant feature with minimum expertise of CRM analysts [34]. Finally, feature subsets obtained by the GMDH algorithm are aggregated to obtain the optimal feature subset through a cost-weighted average approach. The empirical study based on customer data from a cigarette company shows that the new algorithm can provide good prediction, especially for the high-value target customers, and it also helps the decision-makers to identify key customer characteristics.

The remainder of this paper is organized as follows: In Sect. 2, we present a brief description of the feature selection and class imbalance issue, transfer learning and GMDH. Section 3 elaborates on the proposed new algorithm TFSG. In Sect. 4, we evaluate the performance of TFSG and compare it with benchmark feature selection methods on a real-world application. Finally, in Sect. 5, we present the study's conclusions.

2 Related works

2.1 Feature selection and class imbalance

Feature selection first originated from the field of pattern recognition. In customer relationship management, feature selection not only helps companies to identify key determinants of customer behavior, but also saves a large amount of computational time and cost. Thus, we have witnessed various applications of feature selection in customer relationship management. For example, Piramuthu [33] presented a new feature selection methodology based on the blurring measure for customer financial credit risk evaluation. Kim et al. [23] proposed a new evolutionary local feature selection algorithm for customer targeting. Tseng and Huang [39] introduced rough set theory (RST) to feature selection for predicting customer purchasing behavior. Huang et al. [17] applied a new multi-objective feature selection approach to churn prediction in telecommunications service field. Chen and Li [8] utilized an SVM classifier combined with conventional statistical LDA, decision tree, rough sets and *F* score approaches as a feature preprocessing step to discriminate between good and bad customers in credit scoring. Oreskiet al. [31] proposed a heuristic for the feature selection process using genetic algorithms. Farquad [12] developed the support vector machine recursive feature elimination (SVM-RFE) algorithm in the customer churn prediction model. All of these studies have shown the benefits of feature selection.

Class imbalance is often encountered in CRM, especially in customer identification. Class imbalance is a very hot topic in the past decade. It is considered to be one of the ten challenging problems in data mining research [45]. Although many solutions for handling class imbalance have been proposed, most of them focus on the classification task. Only several works attempted to tackle this issue for feature selection task. In the study of Van Hulse et al. [40], the influence of resampling techniques on the feature selection was examined by using five imbalanced microarray expression datasets. Four sampling methods were considered in their work, namely random undersampling, random oversampling, synthetic minority oversampling technique (SMOTE) and Wilson editing. Khoshgoftaar et al. [22] presented a process involving a random undersampling technique for addressing uneven class distribution to select important features in software engineering. Gao et al. [13] proposed a new technique, called SelectRUSBoost, which is a form of ensemble learning that incorporates random undersampling into feature selection to alleviate class imbalance. Yin et al. [46] presented a novel feature selection approach based on class decomposition to cope with the class imbalance issue. As we can see, most of the attempts do not add any new information to the imbalance data and the model performance could be hindered by the absolute rarity.

2.2 Transfer learning

The concept of transfer learning comes from psychology [35], and it illustrates the psychological observation that humans can often benefit from previous learning experiences when learning a new related task. For instance, a good checker player may find it easier to learn playing chess than other people. Thrun [38], Baxter [2] and Caruana [6] introduced the idea into the data mining and machine learning field in the mid-1990s under the labels of "learning how to learn," "lifelong learning" and "multitask learning." At the beginning of this century, Ben-David and Schuller [3] presented the notion of relatedness between learning tasks, which provides theoretical justification for transfer learning. The outline of transfer learning is given in Fig. 1, and a comprehensive survey of transfer learning is provided by Pan and Yang [30].

Instance transfer is the most popular method for transfer learning. It assumes that some instances in the source domains are helpful for training the target domain model, while others could do harm to the target learning task.

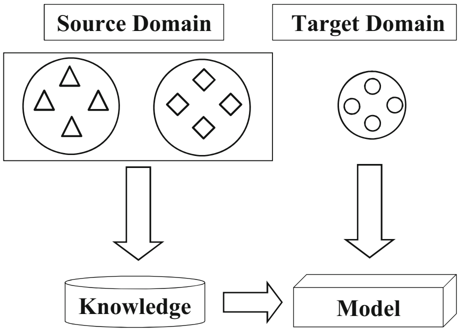

Fig. 1 Outline of transfer learning

Instance transfer uses reweighting or resampling techniques to select those useful source domain instances. Many instance-based transfer learning approaches for classification have already been proposed. For example, Dai et al. [9] proposed a boosting-style method to utilize the old data in related domains. Jiang and Zhai [20] presented an instance-weighting approach to exploit auxiliary information from the source domain based on the difference between the conditional probabilities. Recently, Kamishima et al. [21] modified the bagging ensemble algorithm to develop an algorithm called transfer bagging (TrBagg). TrBagg augments the model performance in the target domain by aggregating the base models that are built on datasets incorporating the source domain data. The proposed approach TFSG is also an instance-based transfer learning method for feature selection. It tries to incorporate information from useful source domain instances to tackle the absolute rarity in the target domain.

2.3 GMDH neural network

The GMDH is an inductive modeling method that constructs a hierarchical (multilayered) network structure to identify complex input–output functional relationships from data. This method was first developed by Ivakhnenko [18] as a multivariate analysis method for complex systems modeling and identification in the 1960s. In the 2000s, Lemke and Muller [26] further developed the GMDH algorithm into the self-organizing data mining algorithm. Since then, GMDH has been successfully applied to many fields such as chemistry [14], engineering [34] and economics [41].

The modeling process of GMDH is analogous to the natural evolution of wheat. It is a self-organizing process based on synthesizing models of increasing complexity and selecting the best solution according to an external criterion. The GMDH first produces some simple elementary models (network neurons) by reference functions and uses them as initial input models at start of modeling process. After generating a large number of competitive models from the initial input models (inheritance), the algorithm selects certain more optimal intermediate models (selection) such that a large number of new competitive models are generated by these intermediate models. This procedure of inheritance and selection is repeated until an optimal complex model is created. According to the optimal complexity theory, as the model complexity increases, the value of external criterion usually decreases first and then reaches a minimum. It later starts to increase again because of model overfitting. The GMDH algorithm will stop when the external performance criterion reaches its minimum, and an optimal complex model is obtained [28].

One advantage of the GMDH is that it can automatically select input features. The modeling process of GMDH is iterative. Starting from simple elementary network neurons produced by reference functions, it generates many intermediate models with different variables. The models that yield better external criteria are selected because they contain inputs that have better ability to explain or predict the output. The input variables from the selected models are used to construct more complex models in subsequent layers. As the modeling process proceeds, the number of variables in the model candidates increases. The procedure will stop when the external criterion reaches its minimum and the optimal complex model with the most effective input variable set is found. In this way, the algorithm can automatically determine the key input feature and exclude the irrelevant ones. Some people have successfully used GMDH for feature selection . For example, El-Alfy and Abdel-Aal [10] presented a technique based on GMDH to identify the most effective content features for classifying emails.

3 Methodology

3.1 Problem definition

Suppose that there is a dataset in the target domain that is composed of m customer instances $D^{\mathrm{T}} = \{(\mathbf{x}_i^T, y_i^T)\}_{i=1}^m$. $\mathbf{x}_i^{\mathrm{T}} \in X$ is an instance in the d-dimensional feature space $X = \{f_1, \ldots, f_d\}$, and it describes customers characteristics. y_i^{T} is the label of the customer class. In this paper, we assume that there are only two types of customers $y_i^{\mathrm{T}} \in \{0, 1\}$, where $y_i^{\mathrm{T}} = 1$ denotes high-value customer and $y_i^{\mathrm{T}} = 0$ represents low-value customer. The goal of feature selection is to find a feature subset $E = \{f_1', \ldots, f_k'\} \subset X, k < d$ that can distinguish high-value customers from low-value ones.

In customer identification, firms often face the situation of class imbalance. We define $P^{\mathrm{T}} \subset D^{\mathrm{T}}$ as the set of high-value customer instances that belong to the minority class, and $N^{\mathrm{T}} \subset D^{\mathrm{T}}$ as the set of low-value customer instances belonging to the majority class. Class imbalance refers to the situation in which $|P^{\mathrm{T}}| \ll |N^{\mathrm{T}}|$. When class imbalance is caused by the relatively small size of the minority class P^{T}, we call it absolute imbalance.

We assume there are n similar customer instances in the source domain, $D^{\mathrm{S}} \in \{(\mathbf{x}_i^{\mathrm{S}}, y_i^{\mathrm{S}})\}_{i=1}^n$, where the feature space $X = \{f_1, \ldots, f_d\}$ is the same as in the target domain. The question now is how we can use customer data D^{T} in the target domain to identify a feature subset E that captures the key customer characteristics in the target domain with the help of auxiliary customer data D^{S} from the source domain.

3.2 Transferred feature selection based on GMDH (TFSG)

We aim to solve the issue of absolute imbalance that is encountered in feature selection and propose a novel instance-based transfer learning technique TFSG. The main idea of TFSG is to import the minority class instances from the source domain to augment the information of the minority class in the target domain by using an ensemble of GMDH. The algorithm flowchart is shown in Fig. 2.

The TFSG algorithm first uses bootstrap sampling on the minority instances of source domain and undersampling on the majority instances of the target domain. In a next step, it combines them with the minority instances of the target domain to produce several training datasets. After that, the GMDH is used on each training dataset to obtain candidate feature subsets, and finally, different feature subsets obtained by GMDH are aggregated with a cost-weighted average approach to obtain an optimal feature ranking. There are three aspects of TFSG that should be stressed. First of all, the minority class dataset P^{S} from the source domain contains both useful and useless instances, so we use the cost-weighted average to eliminate the interference of useless data. More specifically, the cost-weighted accuracy of each base model is computed on the original target training dataset D^{T}. If the value of the cost-weighted accuracy (CWA) is high, then we assume that most of the source domain instances that have been used to train these base models are useful for predicting the customer class in the target domain, and these base models will be assigned larger weights. At the same time, smaller weights are assigned to those base models that have low values of CWA. The final feature ranking score is determined by the weight average. Second, we use GMDH to build the base model, because it can select the feature subset with little expertise. People only need to determine the number of model candidates selected in each layer, which is a very easy task for nearly all problems. Third, one advantage of

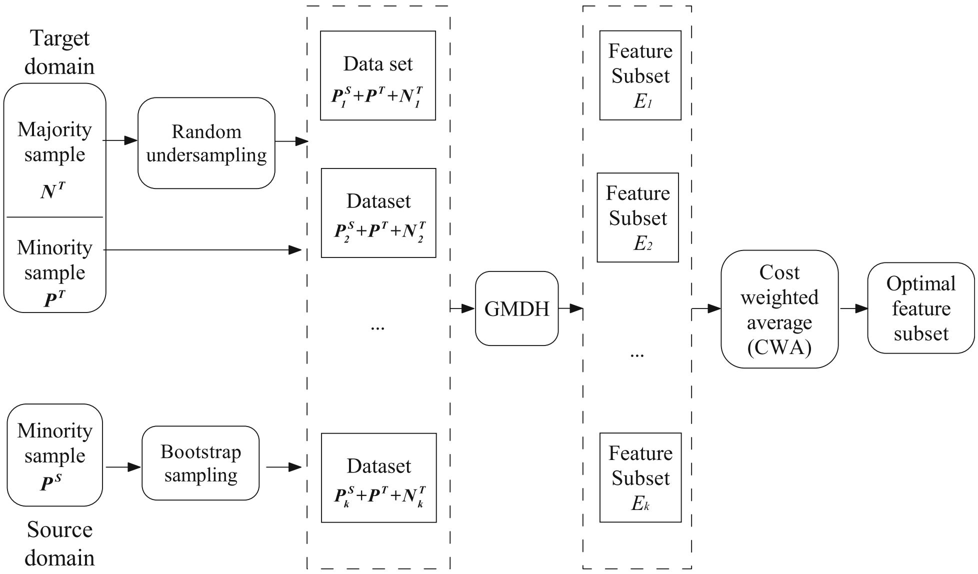

Fig. 2 Flowchart of the TFSG algorithm

ensemble feature selection [1] is stability. Previous research has shown that feature selection is not stable on imbalanced datasets [42], and so we adopt an ensemble of feature selection to make the results more stable. The main steps of the algorithm are as follows:

Step 1: Employ bootstrap sampling on P^{S} to obtain dataset P_i^{S};

Step 2: Use random undersampling on N^{T} to obtain N_i^{T}, where $|N_i^{\mathrm{T}}| = |P_i^{\mathrm{S}}| + |P^{\mathrm{T}}|$;

Step 3: Combine P_i^{S}, P^{T} and N_i^{T} to obtain a training dataset $D_i = P_i^{\mathrm{S}} \cup P^{\mathrm{T}} \cup N_i^{\mathrm{T}}$, and use GMDH to build a base model and obtain the feature subset E_i;

Step 4: Calculate the value of the cost-weighted accuracy (CWA) on the original target dataset D^{T} for each base model built by GMDH as follows:

$$\mathrm{CWA} = (C_N * \mathrm{TNR} + C_P * \mathrm{TPR})/(C_N + C_P) \tag{1}$$

where C_N and C_P are the misclassification cost of the majority and minority class, respectively. TNR and TPR are the true negative rate and true positive rate, respectively;

Step 5: Calculate the feature ranking score q_{ij} for each variable f_j according to the value of CWA as follows:

$$q_{ij} = \begin{cases} \mathrm{CWA} & \text{if} \quad f_j \in E_i \\ 0 & \text{if} \quad f_j \in E_i \end{cases} \quad (j = 1, \ldots, d) \tag{2}$$

Step 6: Repeat step 2 to step 5 to obtain k feature subsets $\{E_1, E_2, E_3, \ldots, E_k\}$ and corresponding ranking scores;

Step 7: Use the following formula to calculate the final aggregated score Q_j for each feature:

$$Q_j = \sum_{i=1}^{k} q_{ij}, \quad j = 1, \ldots, d \tag{3}$$

In the TFSG algorithm, the following GMDH algorithm is used to construct the base models in step 3:

Step 1: Split dataset D_i into two disjoint subsets of the same size $D_i = B \cup C$;

Step 2: Combine input variables in pairs $(f_i, f_j), 1 \le i, j \le d$ and generate model candidates from each combination by using the following quadratic polynomial:

$$y = c_0 + c_1 f_i + c_2 f_j + c_3 f_i \cdot f_j + c_4 f_i^2 + c_5 f_j^2 \tag{4}$$

where $c_0, c_1, \ldots, c_5$ are parameters estimated by ordinary least square (OLS) method (see Fig. 3).

Step 3: Evaluate the external criterion of each model by using the symmetric regularity criterion (SRC) as follows:

$$\mathrm{SRC} = \left(\sum_{i \in B} \left(y_i - \hat{y}_i^C\right)^2 + \sum_{i \in C} \left(y_i - \hat{y}_i^B\right)^2 \right) \tag{5}$$

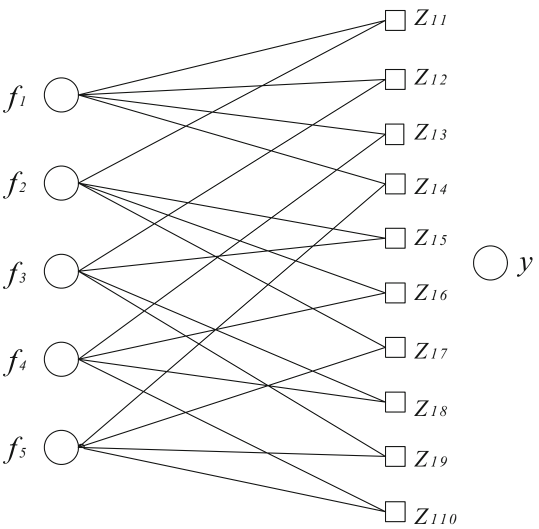

Fig. 3 Generation of candidate models in the first layer

where y_i is the actual output, and $\hat{y}_i^B$ and $\hat{y}_i^C$ are the estimated outputs of the models that are constructed on dataset B and dataset C. Record the minimum of the external criterion R_l from the current layer;

Step 4: Select the F_l best models with the lower criterion values, and use their outputs z_{lt} as new features for the second layer of the GMDH (see Fig. 4).

Step 5: Repeat step 2 to step 4 to produce model candidates for the second layer, the third layer, and so on, until the lowest value of the external criterion at the current layer R_l is greater than that in the previous layer. The model with the minimum external criterion at the $l-1$ layer is then selected as the final model. Figure 5 provides an example where the optimal complex model is obtained at the third layer.

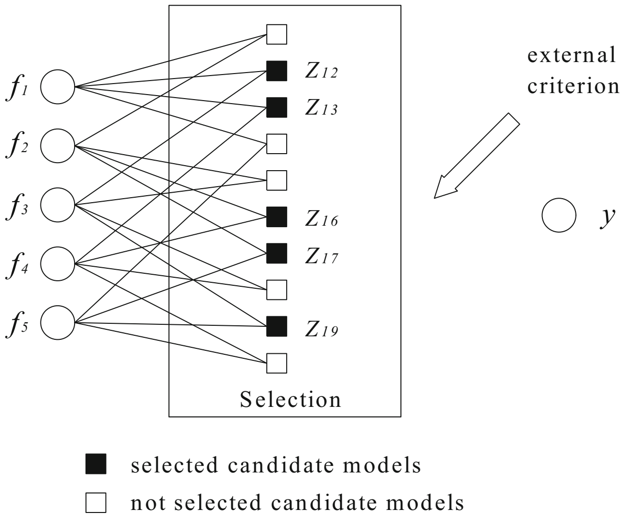

Fig. 4 Selection of candidate models

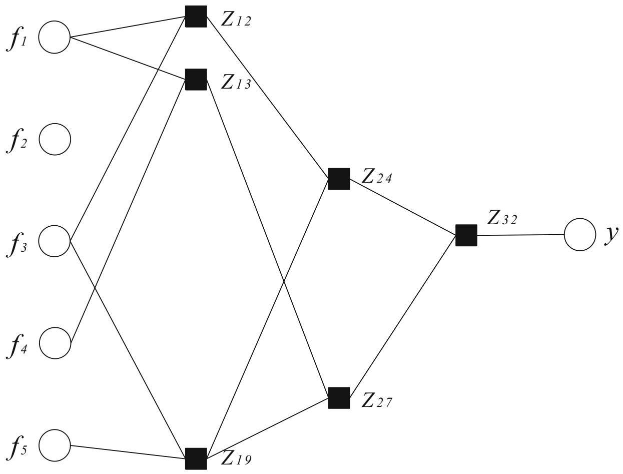

Fig. 5 Generation of the optimal complex model

4 Experiments

In this section, we will evaluate the effectiveness of the new method through experiments on a real-world application.

4.1 Experimental setting

The data were provided by a cigarette company from the Sichuan Province of China. The company intended to launch a new product in the mid-priced cigarette market. Therefore, the managers hoped to understand the drivers of customer behavior in this market segment, and questionnaire surveys were conducted in the Sichuan Province in July 2010. The customers were required to provide socio-demographic background, perceptions or attitudes toward products and behavioral information. According to the amount of smoking, the customers were segmented into heavy (more than 15 cigarettes per day) and mild smokers (less than 15 per day). The heavy smokers are high-value customers. The company obtained a sample with 540 respondents from the city SL (target domain), which is a typical dataset with absolute class imbalance. There are only 138 heavy smokers, which not only has a small size but is also outnumbered by the mild smokers (302 respondents). Fortunately, a similar survey was conducted in the city JY (source domain) several months ago, which yielded 420 respondents with 144 heavy smokers. After eliminating variables with too many missing values, we obtained 57 variables from the survey data. Obviously, it is not convenient to describe the customers and develop marketing program with all of the 57 feature. Thus, the procedure of feature selection is required.

To demonstrate the effectiveness of the proposed method, five feature selection methods are used as benchmarks: ReliefF, information gain (IG), sequential backward selection (SBS), genetic search (GS) and SVM-RFE. ReliefF and IG are filter methods. ReliefF is an extension of the original Relief algorithm [25]. Previous research has shown that it has good performance regardless of the specific data [5]. IG is an information theoretical method that measures the information gain of the class variable when the attribute variable is given. SBS and GS are wrapper methods. SBS starts from a full set and sequentially removes the feature that results in the smallest decrease in the value of the objective function. The genetic search performs the feature search by using the well-known genetic algorithm. SVM-RFE is an embedded method, and it performs feature selection by iteratively training an SVM classifier using linear kernel with the current set of features and removing the least important feature as indicated by the SVM [15].

In our study, the target data (D^{T}) from city SL were randomly divided into a training sample with 70 % of the observations, and a test sample with 30 % of the observations. The data division was performed in a stratified manner to ensure that the proportions of heavy smokers were the same in both training dataset and testing dataset. All methods were applied to the training set to select the feature subsets. For each selected feature subset, a multilayered feedforward neural network was trained on the selected features and then tested on the corresponding test set. We used a multilayered feedforward neural network as the classifier because it is a widely used method in customer classification. In the two wrapper methods (SBS and GS), the multilayered feedforward neural network was also used as the learning scheme. To avoid bias, the above procedures were repeated 30 times, and the average results are reported. Four metrics were used to evaluate experimental results: accuracy, area under the receiver operating characteristic curve (AUC), sensitivity and specificity. All of the four metrics give us more insights into model performance [16] in a class imbalance setting. After discussion with the business experts in the cigarette company, we set the misclassification cost to $C_P = 1$ and $C_N = 4$ because the average profit from heavy smokers is usually four times that of mild smokers in the middle-price cigarette market.

In the experiments, we investigated the performance of the benchmark feature selection methods in different scenarios and compared TFSG with them. Specifically, we attempt to answer the following three questions:

1. How does TFSG perform in comparison with the benchmark methods that are trained on the original target data?
2. Will resampling methods have comparable performance as TFSG in addressing class imbalance?
3. Is there any performance difference between TFSG and the benchmark methods trained on the dataset formed by the direct combination of target and source data?

To answer these questions, we considered three scenarios in which the benchmark methods are built on different datasets. In the first scenario, we only selected the features with benchmark methods on the original target training data D^{T}. In the second scenario, the well-known oversampling method synthetic minority oversampling technique (SMOTE) [7] was used on the target training data D^{T} to generate artificial minority instances and get balanced datasets. SMOTE is a powerful method that has shown good performance in various applications and is still one of the state-of-the-art resampling methods. In the third scenario, we combined the minority class data P^{S} from the source domain with the target data D^{T} without any adjustment, and trained models on the combined dataset $D^{\mathrm{T}} \cup P^{\mathrm{S}}$. All of the benchmark methods were implemented by using the Weka data mining tool with default settings [44]. The TFSG method was programmed by us in MATLAB.

4.2 Experimental results

In this subsection, we will discuss the experimental results to evaluate the effectiveness of the new algorithm. Tables 1, 2 and 3 report the results of TFSG and benchmark methods in the three experimental scenarios. The best values of each metric are highlighted in bold. We will analyze experimental results in the three above-mentioned scenarios, respectively.

Table 1 Experimental results on original target data

Method	AUC	Sensitivity	Specificity	Accuracy
ReliefF	0.6454	0.4556	0.7183	0.6510
IG	0.6982	0.0	**1.0**	**0.7440**
SBS	0.6759	0.2730	0.8904	0.7323
GS	0.6713	0.1821	0.8722	0.6955
SVM-RFE	0.7091	0.4554	0.8217	0.7279
TFSG	**0.7214***	**0.6367***	0.7753*	0.7398

* Statistically significant difference between TFSG algorithm and the best benchmarks method at the 5 % significance level

Table 2 Experimental results on target data using SMOTE sampling

Method	AUC	Sensitivity	Specificity	Accuracy
ReliefF	0.6434	0.5089	0.7049	0.6547
IG	0.6643	0.6363	0.6904	0.6765
SBS	0.7212	0.5456	**0.8416**	**0.7658**
GS	0.6998	0.4732	0.8061	0.7208
SVM-RFE	**0.7246**	0.5552	0.8113	0.7457
TFSG	0.7214	**0.6367**	0.7753*	0.7547

* Statistically significant difference between TFSG algorithm and the best benchmarks method at the 5 % significance level

Table 3 Experimental results on combined dataset

Method	AUC	Sensitivity	Specificity	Accuracy
ReliefF	0.6658	0.4673	0.7417	0.6714
IG	0.6212	0.3121	**0.9234**	**0.7694**
SBS	0.6323	0.3649	0.7696	0.7255
GS	0.6755	0.5449	0.7952	0.7311
SVM-RFE	0.6921	0.3644	0.7473	0.6939
TFSG	**0.7214***	**0.6367***	0.7753*	0.7547

* Statistically significant difference between TFSG algorithm and the best benchmarks method at the 5 % significance level

4.2.1 *Comparison with benchmark methods on original target data*

We first check the performance of the benchmark methods built on the original target data. Table 1 shows the experimental results in this scenario. As Table 1 shows, each benchmark method gives high accuracy and low sensitivity. The sensitivity values of the five benchmark methods are smaller than 0.5, and IG even has a zero value. This finding indicates that the benchmark methods are significantly affected by the uneven class distribution. They cannot manage to identify the minority target customers effectively. Meanwhile, TFSG performs well not only in terms of AUC and accuracy, but also in terms of sensitivity. TFSG ranks first with respect to AUC and sensitivity, and the superiority is indicated by the corrected resampled *t* test. In addition, its accuracy ranks in the second place. Thus, we can conclude that TFSG has better performance than the benchmark methods which are significantly impacted by the class imbalance.

4.2.2 *Comparison with resampling solutions*

Next, we will analyze the experimental results using the SMOTE sampling technique. Table 2 presents the results in that scenario. We can see from Table 2 that the usage of SMOTE helps the benchmark methods improve their ability of identifying the minority class. Note that the degree of improvement is different for different methods. For example, the value of sensitivity increases sharply from 0.0 to 0.6363 for IG after SMOTE sampling, while ReliefF only has a slight increase (from 0.4556 and 0.5089). However, we can find that accuracy and sensitivity of the benchmark methods after using SMOTE sampling cannot catch up with TFSG. TFSG performs the best in terms of sensitivity, and its AUC value ranks second just behind SVM-RFE. In terms of accuracy, it also takes the second place. The analysis shows that resampling can alleviate the influence of class imbalance on the benchmark methods, but TFSG has better performance than the benchmark methods with resampling solutions, especially in terms of accuracy on the minority class.

4.2.3 *Comparison with benchmark methods on combined dataset*

We also compare the performance of the benchmark methods on the dataset that is formed by direct combination of the source and target data. The experimental results are presented in Table 3. As the table shows, the introduction of minority instances from the source data has a mixed influence on the benchmark methods. Some benchmark methods now have a better sensitivity to some extent. For example, the sensitivity of SBS increases from 0.2730 to 0.3649. On the other hand, some methods perform worse after the source data are introduced, such as SVM-RFE whose sensitivity decreases from 0.4554 to 0.3644. So simple combination is inappropriate because the source and target domain have different distributions. Meanwhile, TFSG makes better use of the information from the source domain. It is ranked in the first position for both AUC and sensitivity. The better performance is also confirmed by the *t* test. With respect to the specificity and accuracy, it takes the second place. The experimental results in the third scenario show that TFSG outperforms the benchmark methods trained on the dataset that was formed by direct combination.

4.2.4 *Model interpretability*

Now we can have a closer look at model interpretability. Figure 6 shows the 11 features that are selected by TFSG. The values of relative importance are calculated from the original feature ranking scores using a linear

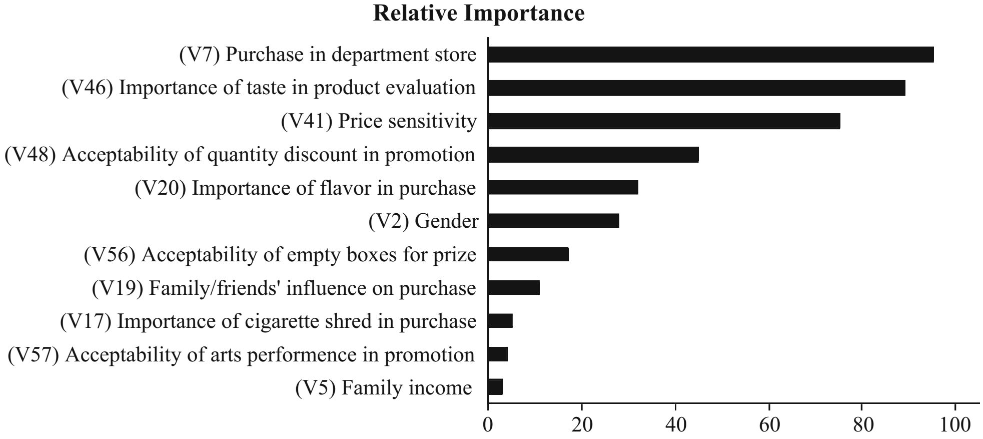

Fig. 6 Relative importance of selected features

Table 4 Experimental results of contingency table analysis

Variable	Value	*df*	*p* value
Purchase in department store	6.687	1	<0.01
Importance level of taste in product evaluation	7.630	3	0.0543
Price sensitivity	5.542	2	0.0626
Acceptability level of quantity discount in promotion	9.943	5	0.0768
Importance of flavor in purchase	8.04	3	0.0453
Gender	3.50	1	0.0614
Acceptability level of empty boxes for prize	12.22	5	0.0320
Family/friend’s influence on purchase	6.414	3	0.0931
Importance of cigarette in purchase	6.624	3	0.0849
Acceptability level of arts performing in promotion	8.264	5	0.1422
Family income	4.203	2	0.1222

transformation, where the largest value will be transformed into 100 and the smallest value will be assigned to 0. We will check the performance of the 11 features one by one to see how they differentiate heavy smokers from mild smokers by using statistical test. Because the 11 selected features are categorical variables, two-way contingency tables are used to evaluate whether a statistical relationship exists between each selected feature and customer type y. The contingency table is a useful tool to analyze the relation between two or more categorical variables [36]. The results are shown in Table 4, where the second column presents the values of the Chi-square statistics, the third column reports degrees of freedom (DF), and the fourth column gives the p values.

From Table 4, we can see that nine out of the eleven selected features yield p values that are smaller than 0.1 based on the Chi-square test. The only two exceptions are *Acceptability level of arts performing in promotion* and *Family income*, which have p values that are close to 0.1. The results show that the null hypothesis that there is no significant relations between the selected features and the customer type is rejected at the 10 % significance level. In other words, there is a significant relationship between the selected features and customer type in the mid-priced cigarette market. In other words, all of the selected features distinguish the heavy and mild smokers effectively and they can be considered to be the key drivers of customer behaviors in the mid-priced cigarette market.

To summarize, the experimental results allow us to answer the three research questions presented in Sect. 4.1: TFSG has better predictive performance than both original benchmark methods and the benchmark methods with resampling solutions. It also makes full use of auxiliary source data given the distribution difference. Hence, it is superior to the benchmark methods trained on the dataset formed by the direct combination of target and source data. In addition, TFSG algorithm results in good model interpretability, and the features selected by it present important features that differentiate different customers.

5 Conclusions

In customer identification, the distribution of low-value and high-value customers is often imbalanced. In this paper, we focus on the issue of class imbalance in feature selection caused by absolute imbalance. A new transferred feature selection algorithm based on the GMDH is proposed to deal with this issue. Based on the idea of transfer learning, the proposed method TFSG helps to select a useful feature subset on the target imbalanced dataset by introducing customer information from related domains. We tested the proposed algorithm on a real-world application to analyze customer characteristics in the cigarette market. It has been demonstrated that the new method TFSG can identify minority high-value customers more effectively than existing feature selection methods. This approach also exhibits better performance than the resampling technique, which is the state-of-the-art method to tackle class imbalance in practice. In addition, the proposed technique can also identify key features that could differentiate high-value and low-value customers with good interpretation. The new algorithm provides a new tool for customer relationship management.

Acknowledgments This work is supported by Natural Science Foundation of China (Grant Nos. 71401115, 71471124, 71172196 and 71572120), MOE (Ministry of Education in China) Youth Project of Humanities and Social Sciences (Grant No. 13YJC630249), Specialized Research Fund for the Doctoral Program of Higher Education of the Ministry of Education of China (Grant No. 20120181120074), and Foundation of Sichuan University (skqy201507). Bing Zhu is supported by postdoctoral fellowships from the China Scholarship Council (CSC).

Appendix

See Table 5.

Table 5 Overview of explanatory variables

Category	ID	Description	Type	Range
Socio-demographic predictor	V1	Age	Continuous	>18
	V2	Gender	Binary	{0, 1}
	V3	Educational level	Ordinal	{1, 2, 3}
	V4	Personal income	Ordinal	{1, 2, 3}
	V5	Family income	Ordinal	{1, 2, 3}
Behavioral predictor	V6	Length of time to smoke the favorite cigarette brand	Continuous	>0
	V7–V12	Whether to buy cigarettes in department store, supermarket, convenience store, grocery, tobacco dealer shop or hotel/restaurant/public entertainment places	Binary	{0, 1}
	V13–V23	Importance level of individual habit, taste, product awareness, package, cigarette shred, tar content, family/friends' influence, flavor, salesman recommendation, product promotion, advertisement in purchase	Ordinal	{1, 2, 3, 4}
	V24–V30	Importance level of price, taste, cigarette shred, individual habit, reputation of production place, tar content, flavor on brand choice	Ordinal	{1, 2, 3, 4}
	V31–V40	Importance level of income change, taste change, product quality, flavor, price, friend/family recommendation, brand popularity, convenience of purchase, attempt of new brand, product promotion in brand switch	Ordinal	{1, 2, 3, 4}
Perception predictor	V41	Price sensitivity	Ordinal	{1, 2, 3}
	V42–V47	Importance level of cigarette shred, cigarette ash, feel, package, taste, tar content in product evaluation	Ordinal	{1, 2, 3, 4}
	V48–V57	Acceptability level of quantity discount, gifts, free samples, TV advertisement, newspaper advertisement, roadside advertisement, poster, SMS lottery, empty boxes for prize, arts performance in promotion	Ordinal	{1, 2, 3, 4, 5, 6}

References

1. Awada W, Khoshgoftaar TM, Dittman D, Wald R, Napolitano A (2012) A review of the stability of feature selection techniques for bioinformatics data. In Proceeding of 13th IEEE international conference on information reuse and integration, pp 356–363
2. Baxter J (1997) A Bayesian/information theoretic model of learning to learn via multiple task sampling. Mach Learn 28(1):7–39
3. Ben-David S, Schuller R (2003) Exploiting task relatedness for multiple task learning. In: Proceedings 16th annual conference on computational learning theory, Washington, DC, USA
4. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
5. Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. Knowl Inf Syst 34(3):483–519
6. Caruana R (1997) Multitask learning. Mach Learn 28:41–75
7. Chawla NV, Bowye KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16(3):321–357
8. Chen FL, Li FC (2010) Combination of feature selection approaches with SVM in credit scoring. Expert Syst Appl 37(7):4902–4909
9. Dai W, Yang Q, Xue G, Yu R (2007) Boosting for transfer learning. In: Proceeding of the 24th international conference on machine learning, ACM Press, pp 193–200
10. El-Alfy EM, Abdel-Aal RE (2011) Using GMDH-based networks for improved spam detection and email feature analysis. Appl Soft Comput 11(1):477–488
11. Faisal A, Gillberg J, Leen G, Peltonen J (2013) Transfer learning using a nonparametric sparse topic model. Neurocomputing 112:124–137
12. Farquad M, Ravi V, Raju S (2014) Churn prediction using comprehensible support vector machine: an analytical CRM application. Appl Soft Comput 19:31–40
13. Gao K, Khoshgoftaar TM, Napolitano A (2012) A hybrid approach to coping with high dimensionality and class imbalance for software defect prediction. In: Proceeding of the 11th international conference on machine learning and applications, pp 281–288
14. Ghanadzadeh H, Ganji M, Fallahi S (2012) Mathematical model of liquid-liquid equilibrium for a ternary system using the GMDH-type neural network and genetic algorithm. Appl Math Model 36(9):4096–4105
15. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422
16. He H, Garcia E (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
17. Huang B, Buckley B, Kechadi TM (2010) Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. Expert Syst Appl 37(5):3638–3646
18. Ivakhnenko AG (1968) The group method of data handling—a rival of the method of stochastic approximation. Sov Autom Control 1–3:43–55
19. Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data Anal 6(5):429–450
20. Jiang J, Zhai C (2007) Instance weighting for domain adaptation in NLP. In: Proceedings of the 45th annual meeting of the Association for Computational Linguistics, pp 264–271
21. Kamishima T, Hamasaki M, Akaho S (2009) TrBagg: a simple transfer learning method and its application to personalization in collaborative tagging. In: Proceeding of ninth IEEE international conference on data mining, pp 219–228
22. Khoshgoftaar TM, Gao K, Seliya N (2010) Attribute selection and imbalanced data: problems in software defect prediction. In: Proceeding of international conference on tools with artificial intelligence, pp 137–144
23. Kim Y, Street W, Russell G, Menczer F (2005) Customer targeting: a neural network approach guided by genetic algorithms. Manag Sci 51(2):264–276
24. Kim Y (2006) Toward a successful CRM: variable selection, sampling, and ensemble. Decis Support Syst 41(2):542–553
25. Kononenkom I (1994) Estimating attributes: analysis and extensions of RELIEF. In: Proceedings of the European conference on machine learning, pp 171–182
26. Lemke F, Mueller J (2003) Self-organising data mining. Syst Anal Model Simul 43(2):231–240
27. Luo Y, Liu T, Tao D, Xu C (2014) Decomposition-based transfer distance metric learning for image classification. IEEE Trans Image Process 23(9):3789–3801
28. Mueller JA, Lemke F (2000) Self-organizing data mining: an intelligent approach to extract knowledge from data. Libri Books, Berlin
29. Ngai E, Xiu L, Chau D (2009) Application of data mining techniques in customer relationship management: a literature review and classification. Expert Syst Appl 36:2592–2602
30. Pan S, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
31. Oreski S, Oreski G (2014) Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Syst Appl 41(4):2052–2064
32. Pan S, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
33. Piramuthu S (1999) Feature selection for financial credit-risk evaluation decisions. INFORMS J Comput 11(3):258–266
34. Sheikholeslami M, Sheykholeslami F, Khoshhal S, Mola-Abasia H, Ganji D, Rokni H (2014) Effect of magnetic field on Cu-water nanofluid heat transfer using GMDH-type neural network. Neural Comput Appl 25:171C178
35. Skinner BF (1953) Science and human behavior. Colliler-Macmillian, New York
36. Smith SM, Albaum GS (2005) Fundamentals of marketing research. Sage, Thousand Oaks
37. Stepashko VS, Yurachkovskiy YP (1986) The present state of the theory of the group method of data handling. Sov J Autom Inf Sci 19(4):36–46
38. Thrun S (1996) Is learning the N-th thing any easier than learning the first? In: Proceedings of NIPS-96, pp 640–646
39. Tseng TL, Huang CC (2007) Rough set-based approach to feature selection in customer relationship management. Omega 35(4):365–383
40. Van Hulse J, Khoshgoftaar TM, Napolitano A, Wald R (2009) Feature selection with high-dimensional imbalanced data. In: Proceeding of the 2009 IEEE international conference on data mining workshops, pp 507–514
41. Venkatesh K, Ravi V, Prinzie A, Van den Poel D (2014) Cash demand forecasting in ATMs by clustering and neural networks. Eur J Oper Res 232(2):383–392
42. Wang H, Khoshgoftaar TM, Napolitano A (2012) An empirical study on the stability of feature selection for imbalanced software engineering data. In: Proceeding of 11th international conference on machine learning and applications (ICMLA), pp 317–323
43. Weiss GM (2004) Mining with rarity: a unifying framework. ACM Sigkdd Explor Newsl 6(1):7–19
44. Witten IH, Frank E (2005) Data mining practical machine learning tools and techniques. Morgan Kaufmann, San Francisco
45. Yang Q, Wu X (2006) 10 challenging problems in data mining research. Int J Inf Technol Decis Mak 5(4):597–604
46. Yin L, Ge Y, Xiao K, Wang X, Quan X (2013) Feature selection for high-dimensional imbalanced data. Neurocomputing 105(1):3–11