International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I11P204

Volume 71 Issue 11, 33-44, November 2023

Original Article

Soft Computing based Dual Way Data Modification to Deal with Data Imbalance Problem: Applied to Churn Prediction in Credit Card Users

M.A.H. Farquad[1], Patlolla Venkat Reddy[2], Mohammad Sanaullah Qaseem[3], Syeda Husna Mehanoor[4]

[1 2]*School of Technology, Woxsen University, Hyderabad, Telangana, India.*
[3]*Department of Computer Science, Nawab Shah Alam Khan College of Engineering and Technology, India.*
[4]*Department of Computer Science, Malla Reddy Engineering College for Women, Telangana, India.*

[2]*Corresponding Author: venkatreddy.patlolla@woxsen.edu.in*

Received: 03 August 2023 Revised: 20 September 2023 Accepted: 20 October 2023 Published: 04 November 2023

Abstract *- The data generated by the industry is imbalanced in nature, with nil or least number of samples about customers who are very important to the business, and the industry cannot take chances of losing them to their competitors. Hence, it becomes highly impossible to understand who is important and who is not. It is also a fact that soft computing algorithms tend to produce sub-optimal solutions using imbalanced training data. This paper proposes a data modification procedure to deal with the data imbalance problem. The proposed approach consists of three major steps, viz. (i) feature ranking, (ii) support vector extraction and vector modification and (iii) prediction. Feature ranking is first employed, and top features are selected for further processing. Support vectors are extracted using SVM, and target values of the extracted SVs are replaced with the predictions of trained SVM models, resulting in SV(P) data. Later, during the prediction step, various classifiers are evaluated. The dataset analyzed in this research study pertains to churn prediction in bank credit card customers, with only 6.76% of the samples representing a churner (shifting loyalties to competitors). The classifier's sensitivity has been accorded the highest priority while evaluating the classification algorithms in this research. It is observed that the soft computing techniques employed in this study outperformed and yielded better sensitivity using the proposed modified SVs(P) data compared to the results obtained using other training data.*

Keywords *- Feature ranking, Data modification, Churn prediction, Class imbalance problem, Support Vector Machine.*

1. Introduction

In recent years, the machine learning community has focused on dealing with class-imbalanced data. The class imbalance problem arises when instances of one of the classes (majority class) are much more than the other class (minority class). The class distribution ratio between the majority and minority classes may be 100:1 or 1000:1. In other words, the instances of majority class outnumber the amount of minority class instances. Usually, the study's objective is to understand and predict minority class instances [1]. Many real-world applications such as medical diagnosis, telecommunications, intrusion detection and fraud detection are recognized as class imbalance problems where the data generated is highly imbalanced in nature.

For a couple of decades, the machine learning community has continued to put efforts into dealing with the data imbalance problem [2], where class overlapping, cost of the misclassified class, small-sized training data and small disjuncts of the imbalanced data sets are points of major concern.

1.1. Issues Related to Imbalanced Data Include

- The cost of a wrong prediction of the minority class is much higher than the cost of a wrong prediction of the majority class.
- Soft computing algorithms provided with no or least number of training samples to learn about minority class instances.
- Soft computing algorithms find it difficult to cope with the huge difference between majority-class instances and minority-class instances.
- It is observed that machine learning algorithms tend to underperform and produce unsatisfactory or sub-optimal models, resulting in no trust in such models by the industry.

It has been observed that researchers have proposed many approaches to resample the imbalanced data to bring balance in the data and to obtain at least enough number of samples pertaining to minority class for classifiers to learn from and predict better about minority class.

This paper proposes a two-way data modification using feature ranking and support vector extraction using SVM. The proposed approach is composed of three major steps: *(i) Feature Ranking, (ii) Support Vector extraction and Data Modification, and (iii) Prediction*. Accuracy, Sensitivity, Specificity and AUC are the performance measures considered while dealing with data imbalance problems. Based on the nature of the study in this research, we accorded high importance to sensitivity (predicting the churner as a churner).

The rest of the paper is organized as follows. Section 2 presents the research reported earlier about dealing with imbalanced data. Later, in Section 3, the proposed data modification and its architecture are presented in detail. Section 4 presents a brief overview of the soft computing techniques employed in this research study. Further, Section 5 presents the information about the dataset used and the experimental setup followed in the current study. The empirical analysis is then presented in Section 6. Finally, Section 7 presents the conclusions drawn based on the yielded empirical results.

2. Literature Review

The most simple and easy procedures proposed to deal with the data imbalance problem include randomly eliminating majority class instances, i.e., under-sampling, randomly duplicating minority class instances, i.e., over-sampling and adjusting misclassification costs.

In the earlier stages of this research, SMOTE (Synthetic Minority Over-sampling TEchnique) has proved to be the most effective and efficient way of dealing with class imbalance problems, where synthetic data instances are generated for minority class instances to match the number of majority class instances resulting in balanced data [3]. Previous studies have mainly applied the Synthetic Minority Over-sampling Technique (SMOTE) [32]. Barendela et al. have proposed a procedure to compensate for an imbalance in data by employing a weighted distance function to be used in the classification phase of k-NN [4].

They concluded that their method does not alter the class distribution of the available data but improves the classifier's sensitivity. Later, hybrid approaches combining two or more approaches were reported, such as SMOTE+TOMEK and SMOTE+ENN [5]. Guo and Viktor have analyzed the efficiency of boosting, including over-sampling and concluded that boosting improves the prediction accuracy of the classifier [6].

The application of clustering for majority and minority class samples was later analyzed. First, the clusters are made from the available data and the clusters in the majority class are over-sampled to balance the class distribution ratio [7]-similarly, the use of k-means-based under-sampling and agglomerative hierarchical clustering-based over-sampling [8]. Borderline SMOTE is one of its kind technique which identifies minority samples at the borderline and generates synthetic data using SMOTE [9]. The SMOTE-Bootstrap hybrid was proposed where SMOTE is applied for over-sampling, and Bootstrap is employed for under-sampling [10]. Fernandez et al. have used 2-tuple genetic tuning to generate a fuzzy rule base and concluded that it enhances the prediction accuracy of the fuzzy rule base classifier systems [11]. Similarly, a support vector machine-based EnSVM classifier is employed to resample the data [12].

Later, Alibeigi et al. explored the efficiency of density-based feature selection (DBFS) to deal with imbalanced data [13]. They have concluded that their proposed approach is suitable when the available data is small and imbalanced in nature. Further, the application of the Radial Basis Function classifier in combination with SMOTE and PSO is proposed [14]. Similarly, the Che-PmRF hybrid data balancing approach, which combines the efficiency of PSO-based sampling, MR-based feature selection and RF base classification, is proposed to deal with class imbalance problems [15].

The majority-weighted minority over-sampling technique (MWMOTE) is proposed to deal with specifically minority class [16]. They employed their proposed approach on almost 20 real-world data sets, and an extensive report has been published. After discovering the noise in the data, researchers have proposed a data balancing technique to deal with imbalanced and noisy data [17]. They employed 7 different sampling techniques and carried out extensive experimentation before drawing any conclusion. D'Addabbo and Rosalia have proposed a hybrid approach including a support vector machine and parallel selective sampling to deal with data imbalance problems and concluded that PSS-SVM outperforms respective stand-alone SVM because of no convergence [18].

In [19], to reduce Bayes error, a Near-Bayesian Support Vector Machine (NBSVM) is proposed. The most effective feature of NBSVM is that it deals with noisy data. Later, a bagging-based ensemble method (which does not affect original class distribution) is proposed to deal with the data imbalance problem [20].

Zhu et al. have proposed a new data balancing approach called class weights random forest where majority and minority class instances with high accuracy are selected and concluded that their proposed approach would yield the best classification results [21]. In recent times, hybrid data

sampling approaches that include ensemble classification techniques have been proposed to deal with class imbalance problems. Tsai et al. have proposed under-sampling using instance selection and clustering, where instances belonging to the majority class are clustered together, and an instance selection method was employed to sort irrelevant instances [22].

A novel hybrid approach called Ant Colony Optimization Resampling (ACOR) was proposed to deal with imbalanced data [23]. ACOR first balances the data using random oversampling and later employs ant colony optimization to get a sub-optimal subset from oversampled data. Further, implications of Decision Tree C5.0 and cost-sensitive learning hybrid were proposed to deal with multiclass imbalanced data [24]. First, the decision tree model was created; later, the minimum cost model was obtained using cost-sensitive learning. It is reported that the proposed hybrid approach yielded better results compared to ID3 and C4.5 algorithms.

A novel hybrid approach using Genetic Algorithm-based error classification is proposed to deal with the class imbalance problem [25]. Error identification in data is carried out using PCA and GA. It is concluded that this proposed hybrid approach enhances the processing time. Further, to overcome the suboptimal solutions obtained using an Extreme Learning Machine (ELM), a novel hybrid combining G-Mean-based cost function is proposed for ELM to obtain optimal solutions when imbalanced data is used [26]. Susan et al. [27] have employed random under-sampling of majority-class instances and oversampling of minority-class instances to bring balance to training data. They employed various intelligent versions of oversampling methods. They reported that the decision tree learns better about classes when balanced data is used for training after implementing the proposed data modification. The majority of prior studies have been categorized into those demonstrating methods to enhance predictive performance through the application of diverse machine learning algorithms [28].

Lalwani et al. employed a Gravitational Search Algorithm (GSA) for feature selection and data dimension reduction, enhancing understanding. They reported that optimized algorithms combined with ensemble learning yielded improved results [29]. Further, future studies will include water purifier behavior data for enhanced churn prediction. Given data size complexity, GSA-like feature engineering is valuable. Later, in the realm of customer churn prediction, introduced innovative and efficient hybrid resampling approaches were introduced, which encompass SMOTE-ENN and SMOTE Tomek-Links methods [30].

A recent development introduced an ensemble learning method that seamlessly integrates clustering and classification algorithms for customer churn prediction [31]. Furthermore, an additional concern arises from a class imbalance within customer churn prediction.

2.1. Motivation for the Proposed Approach

The customer-related data generated by the industry is huge in size and imbalanced in nature, with very little or nil information about the customers who are important. It becomes challenging for the industry to identify and retain the most valuable customers. Learning in soft computing techniques is biased towards majority class instances, and learning very little or completely ignores the minority class instances (a major concern). It has become a challenge for the soft computing fraternity to come up with solutions to deal with imbalanced data problems. Hence, in this research, we have proposed a data modification approach to balance the data, resulting in better training of the prediction algorithm.

Feature ranking is one of the most employed procedures to rank available features and ignore the features with the least impact on the target variables. Unnecessary features lead to over-training of the algorithm and end up giving suboptimal or unreal results. Hence, feature ranking is first employed in the proposed approach only to consider top-ranked features to train the classification model.

SVM has proved to be one of the best soft computing techniques introduced to deal with binary classification problems. SVM first tries to find the maximum separating boundary between the classes and considers the characteristics of the instances falling on the boundary (called Support Vectors) for prediction purposes. It is observed that SVM tends to give better results by using only a subset of the training data, i.e. SVs. Hence, the current study employs SVM to reduce the number of instances for training classification models.

In this study, we propose a data modification procedure to balance the data without losing the very nature of the problem and improve the efficiency of classifier learning. We first rank the features; later, extracted support vectors are modified, and then trained classifiers using modified training sets. In this study, "*Churn Prediction in Bank Credit Card Customers*", data is analyzed, which is medium-sized (14814 instances) and unbalanced in nature (93% non-Churners:7% Churners).

3. Proposed Data Modification Procedure

In this paper, we proposed a two-way data modification procedure to deal with the data imbalance problem, and the proposed approach is evaluated using Churn prediction data related to bank credit card customers. The proposed data modification approach is composed of three major steps. *(i) Feature ranking, (ii) Support vector extraction and vector modification and (iii) Prediction*. The complete architecture of the proposed research is presented in Figure 1.

3.1. Step 1: Feature Ranking

Feature ranking is employed using *Chi-Square Statistics, Gini Index, PCA and SVM-RFE*. This results in the reduction of dataset size horizontally. Table 2 shows the relevancy of the attributes given by the techniques employed for feature selection. We considered one-third of the total number of available attributes as important, i.e., the top 7 attributes for the next step of the simulation. It should be noted that we have not followed any specific procedure to decide on the number of attributes to consider as the most relevant feature for further experimentation.

3.2. Step 2: Support Vector Extraction and Vector Modification

During this step, SVM is employed as a classifier using full-featured data and reduced-featured data and Support Vectors (SVs) are extracted (1 and 2 in step 2). Later, actual target values of the extracted SVs are replaced by the predictions of trained SVM, respectively (3 in step 2). Resulting in a change in the dataset, with the subset of the data in the form of SVs with SVM-predicted corresponding target values (*SVs(P)*) (4 in step 2). It is worth mentioning that when reduced feature data is used, the modified *SVs(P)* is the data set that is reduced horizontally as well as vertically. This provides a simpler training set for predictive modeling that generates less complex classification models without compromising the prediction accuracy of the classifier.

3.3. Step 3: Prediction

Naïve Bayes, Decision Tree, Deep Learning and Logistic Regression classifiers are employed for prediction purposes. The classification techniques analyzed in this study are either simple or best-performing. Conclusions are drawn after extensive experimentation using various sets of training data available and modified data, as mentioned below. Performance measure Sensitivity is accorded high priority for the churn prediction problem analyzed in this study.

4. Brief Overview of the Technique Employed

4.1. Feature Ranking

Feature ranking is the process of ranking the available number of features in order of relevance based on the relation between the input attributes and the target variable. Feature ranking can be employed using both statistical and soft computing techniques. We have employed Chi Squared Statistics, Gini Index, PCA and SVM-RFE for extensive analysis for feature ranking purposes.

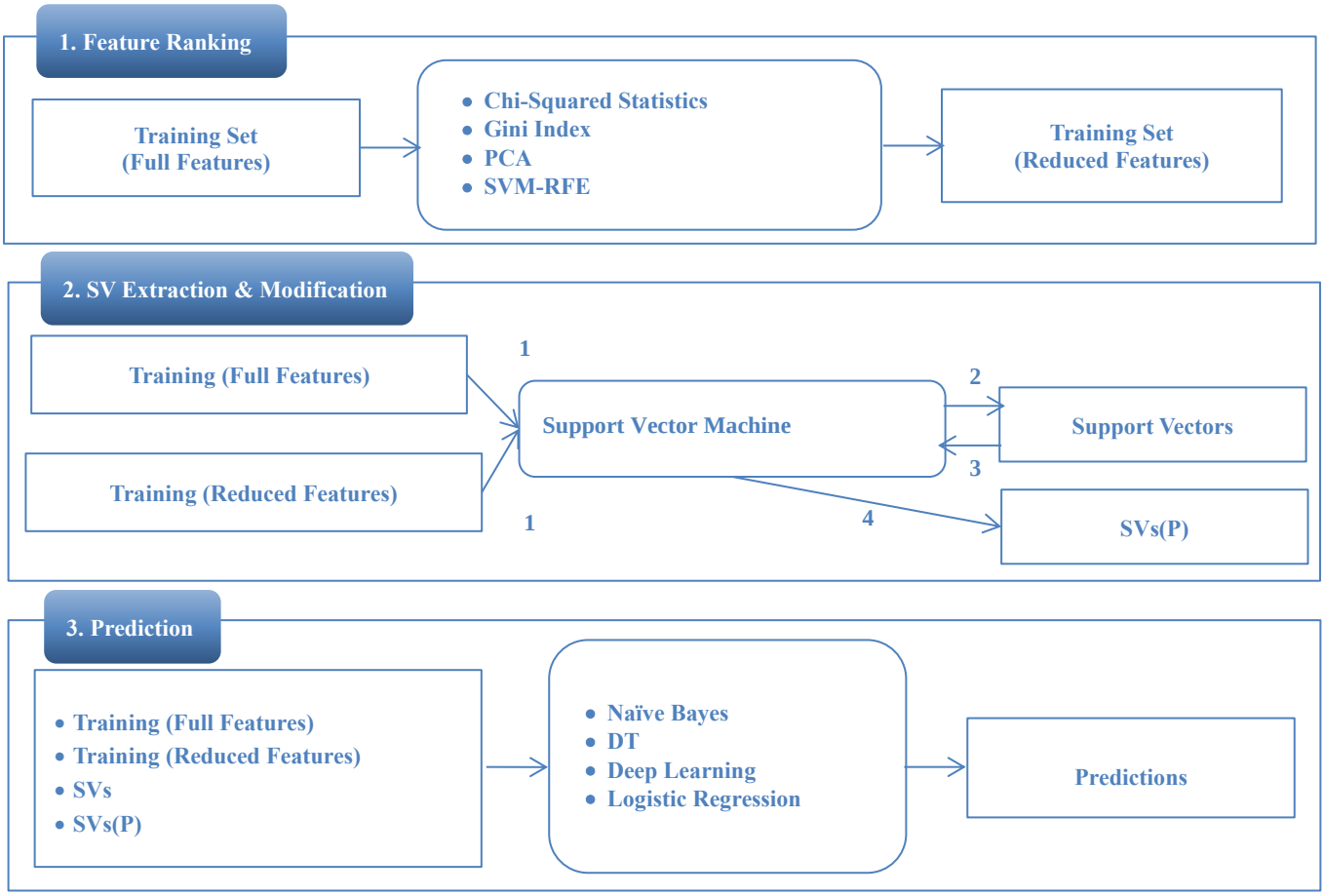

Fig. 1 Proposed data modification architecture and classification

4.1.1. Chi Squared Statistics

Chi Squared Statistics is employed to calculate the impact of the attributes with respect to the target variable. Likewise, the higher the impact, the more relevant the attribute is. It is a nonparametric approach that calculates the difference between the observed frequency and the expected frequency. Because of the nominal nature of the data used, chi-square statistics consider frequencies (*observed and expected*) instead of *mean and variance*. The value of the chi-square statistic *Ch* is given by;

$$Ch = \sum \left[\frac{(O - E)^2}{E} \right]$$

Where *O* is the observed frequency, and *E* is the expected frequency.

4.1.2. Gini Index

The Gini index is employed to calculate the impurity in the given data and is the most employed measurement of equality. Gini index values are yielded by attributes and their impact on the target attribute by class distributions if the available data is split using the given attribute.

4.1.3. Principal Component Analysis

The primary objective of the principal component analysis is to reduce the number of attributes based on their correlation with each other. PCA reduces the number of attributes without losing the variation present in the data, also called Principal Components or simply PCs. Variation in the components decreases as we go down the order, i.e., the first PC retains the maximum variation present in the original components. Scaled data should be provided as the results of PCA are sensitive to the relative scaling. Correlation depicts the relation between two attributes ranging from -1 to +1.

4.1.4. SVM-Recursive Feature Elimination

The SVM-RFE algorithm is employed for feature selection in the present study. The sequential backward elimination procedure is applied over the nested subset of features, which considers all the features available in the data and eliminates one feature at a time. At every step, the weight vector w of a linear SVM is applied to calculate the ranking score for features. Likewise, the i^{th} feature with the least ranking score.

$$c = (w)^2$$

is eliminated, where w is the corresponding weight vector.

In other words, the feature removed using ranking criterion, i.e. $c = (w)^{2,}$ changes the objective function the least. The objective function J chosen in SVM-RFE is,

$$J = \frac{1}{2} \|w\|^2$$

4.2. Predictive Modeling

The prediction efficiency of the classifier is evaluated during this step. We have chosen *Naïve Bayes Classifier, Decision Tree, Deep Learning and Logistic Regression* for classification purposes and to evaluate the proposed data modification procedure. The Naïve Bayes Classifier is based on the probability of the target value depending on input values. DT classifiers generate a decision tree. Deep learning is employed as it is reported to be best performing in recent times. LR is observed to be the simplest classification model. A brief overview of the classification approaches evaluated in this study is given below.

4.2.1. Naïve Bayes Classifier

The Naive Bayes classifier is from the probabilistic classifier family, which applies the Bayes theorem with naive assumptions over the attributes. The model generated by the Naïve Bayes classifier is computationally simple and naïve. Naïve Bayes have demonstrated better prediction accuracy and considered consuming less time when dealing with large-scale and imbalanced datasets. Hence, this research selects the Naïve Bayes classifier for evaluation purposes.

Naïve Bayes classifier is inspired by the Bayes Theorem

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

Where, *X* (input variables) and *y* (output variable) and $P(y \mid X)$ is the probability of *y* given input features *X*. As *P(X)* is constant, proportionality is introduced.

$$P(y|X) \propto P(y) * \prod_{i=1}^{n} P(x_i|y)$$

The goal of Naïve Bayes is to choose the class *y* with the maximum probability. *Argmax* is simply an operation that finds the argument that gives the maximum value from a target function.

$$y = argmax_y \left[P(y) * \prod_{i=1}^{n} P(x_i|y) \right]$$

4.2.2. Decision Tree

Decision tree is one of the simplest and one of the most employed soft computing algorithms, which mimics human-level thinking. Each node in the decision tree represents an attribute, the link represents a rule, and every leaf represents the outcome. The basic objective of generating a decision tree using given training data is to minimize the error at the leaf level. The path reaching each leaf gives an *if-then* form of understanding to the user. At times, pruning is applied when leaf nodes do not add to the discriminative power of the decision tree, which helps in generating general trees and avoids generating over-specific and over-fitted trees.

4.2.3. Deep Learning

Deep learning applies neural network architecture and is often referred to as deep neural networks. It is a multi-layered feed-forward neural network trained using a back propagation method. Deep learning involves an artificial neural network with more than one hidden layer, consisting of neurons with various activation functions such as *tanh*, *rectifier* and *maxout*. Adaptive learning rate, momentum training and level 1 or level 2 regularization enable deep learning to have better predictions.

4.2.4. Logistic Regression

Logistic Regression accepts real-valued inputs and predicts based on the probability of the input belonging to the default class. Threshold values can be set for the probabilities, such as if the probability is >0.5, the output is class 0; otherwise, the output is class 1. Below is an example logistic regression equation:

$$y = \frac{e^{(b_0 + b_1 * x)}}{(1 + e^{(b_0 + b_1 * x)})}$$

Where y is the predicted output, b0 is the bias or intercept term, and b1 is the coefficient for the single input value x. Each column in the input data has an associated b coefficient (a constant real value) that must be learned from given training data.

5. Methodology

5.1. Dataset Description

The dataset analyzed in this study consists of the information of the customers from a Latin American Bank that suffered from an increasing number of churns. Attribute information of the dataset used in this study is presented in Table 1. These attributes include sociodemographic and behavioral data about each customer.

The dataset consists of a total of 14814 records, of which 93.24% (i.e., 13812 of 14814) of the records represent good customers, whereas only 6.67% (1002 of 14814) of the records represent churned customers (Business Intelligence Cup 2004) [33]. Class distribution makes the data set imbalanced in terms of churners versus non-churners and medium scale in terms of the number of instances.

5.2. Experimental Setup

In this research, we have carried out the experiments using a novel evaluation strategy where, prior to model building, 20% of the original data is taken out and called a *validation set* that is used to validate the final prediction model. This is done to evaluate the proposed approach in real-time, as the validation set is never used for training or test purposes. The remaining 80% of the data is then used for training and testing purposes using the holdout method 80:20.

Table 1. Attribute information

Attribute #	Description
1	Credit in the current month
2	Credit in the previous month
3	Credit in before last month
4	Number of credit card transactions in the current month
5	Number of credit card transactions in the previous month
6	Number of credit card transactions before last month
7	Salary or Income
8	Educational level
9	Age
10	Gender
11	Civilian status
12	Number of web transactions in the current month
13	Number of web transactions in the previous month
14	Number of web transactions before last month
15	Margin for the company in the current month
16	Margin for the company in the previous month
17	Margin for the company in last before the month
18	Margin for the company before 3 months
19	Margin for the company before 4 months
20	Margin for the company before 5 months
21	Margin for the company before 6 months
22	Class Variable (Non-Churner; Churner)

Table 2. Attribute weights in descending order using available training data

Weight Order	Chi-Square	Gini Index	PCA	SVM-RFE
1	12	1	7	14
2	7	2	3	4
3	14	3	1	11
4	6	4	2	5
5	5	5	15	1
6	4	6	17	12
7	13	15	18	10
8	18	16	21	13
9	19	12	9	21

As the available test plays its role in finalizing the parameter values for the predictive modeling, we have drawn our conclusions based on the results yielded against the validation set.

The datasets used for training the classifiers are as follows.

1. Training data with full features
2. SVs with full features
3. *SVs(P)* with full features (Proposed)
4. Training data with reduced features
5. SVs with reduced features
6. *SVs(P)* with reduced features (Proposed)

6. Empirical Analysis

All the experiments are carried out using the RapidMiner data mining tool [34, 35] (available free for education and research purposes). The original class distribution ratio is maintained while preparing validation data and training data. Business experts find it essential to find potential churners in time. Hence, *sensitivity* is accorded the highest priority, and conclusions are drawn.

Sensitivity is the measure of the proportion of the true positives (churner) which are correctly identified.

$$\text{Sensitivity} = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

Specificity is the measure of the proportion of the true negatives (good) which are correctly identified.

$$\text{Specificity} = \frac{True\ Negative}{(True\ Negative + False\ Positive\ )}$$

Accuracy is the measure of the proportion of TP and TN that are correctly identified.

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{(True\ Positive + True\ Negative + False\ Positive + False\ Negative)}$$

In this study, we have employed Chi Squared Statistics, Gini Index, PCA and SVM-RFE for feature ranking purposes.

Table 3. Results yielded by SVM while support vector extraction

# of Features	Sensitivity	Specificity	Accuracy
Full Features	45	94.9	91.53
Chi Square	41.5	74.2	71.81
Gini Index	72.5	85.01	84.17
PCA	38.5	92.19	88.49
SVM-RFE	64.71	63.45	63.94

Table 4. SVs Class distribution

Feature Selection	Total	Churner	Good	Distribution
Chi-Square	1524	144	1380	9:91
Gini Index	902	78	824	9:91
PCA	1513	126	1387	9:91
SVM-RFE	624	68	556	11:89

Table 5. Vs(P) Class distribution

Feature Selection	Total	Churner	Good	Distribution
Chi-Square	1524	702	822	46:54
Gini Index	902	423	479	47:53
PCA	1513	643	870	43:57
SVM-RFE	624	346	278	55:45

Later, we considered the first 7 attributes to be considered relevant and further experiments were carried out. Table 2 presents the ranks yielded by various feature ranking techniques. Table 3 presents the sensitivity yielded by SVM using full feature data and reduced feature data. The SVM models created at this point are then used for obtaining predictions for the extracted SVs. The actual target values of SVs are replaced by the predicted target values to obtain *SVs(P)* (the proposed modified data).

Tables 4 and 5 present the class distribution ratios of extracted SVs and modified *SVs(P)*. It is observed that extracted SVs present a similar class distribution as it is available in the original training data. At the same time, class distribution is observed to be more balanced after the proposed data modification procedure, i.e., *SVs(P)*. SVM tries to find an optimal separating hyperplane in such a way that it correctly classifies most or all instances and separates the instances of classes as far as possible by minimizing the risk of misclassification. *C* (acceptable misclassification error) and kernel are two major parameters to be considered while developing the SVM model.

It is observed that with a high value of C, SVM tends to misclassify majority class instances, giving away more minority class instances. Hence, the modified data has an additional number of minority class instances that also balance the class distribution ratio without compromising on the sensitivity of the SVM model. It should be noted that the majority class instances that are predicted as minority class instances are not random in nature but the instances which are analogous to minority class instances.

Table 6. Naïve Bayes classification using full training data

# of Features	Sensitivity	Specificity	Accuracy
Full	70.5	75.05	74.75
Chi Square	65.5	82.59	81.43
Gini Index	69	80.12	79.37
PCA	8.5	93.34	87.61
SVM-RFE	69	84	82.98

Table 7. Naïve Bayes classification using SVs only

# of Features	Sensitivity	Specificity	Accuracy
Full	82	38.99	41.9
Chi Square	3.5	90.01	84.17
Gini Index	3	94.57	88.39
PCA	83	44.97	47.54
SVM-RFE	0	100	93.25

Table 8. Naïve Bayes classification using *SVs(P)*

# of Features	Sensitivity	Specificity	Accuracy
Full	9	75.81	71.3
Chi Square	70.5	74.73	74.44
Gini Index	65.5	74.08	73.5
PCA	11.5	66.98	63.23
SVM-RFE	**96**	16.47	21.84

Table 9. Decision tree Classification using full training data

# of Features	Sensitivity	Specificity	Accuracy
Full	42.5	95.98	92.37
Chi Square	0.5	99.06	92.4
Gini Index	2	99.89	93.28
PCA	52	95.69	92.74
SVM-RFE	0	100	93.25

Table 10. Decision tree classification using SVs only

# of Features	Sensitivity	Specificity	Accuracy
Full	0.5	99.49	92.81
Chi Square	1	97.25	90.75
Gini Index	0	99.28	92.57
PCA	1.5	98.55	92
SVM-RFE	0	100	93.25

Table 11. Decision Tree classification using *SVs(P)*

# of Features	Sensitivity	Specificity	Accuracy
Full	50	86.31	83.86
Chi Square	44	58.69	57.7
Gini Index	74	65.93	66.48
PCA	56	63.5	63
SVM-RFE	74.5	55.79	57.06

Table 12. Deep learning classification using full training data

# of Features	Sensitivity	Specificity	Accuracy
Full	30.5	97.79	93.25
Chi Square	0	99.96	93.21
Gini Index	53.5	97.94	94.94
PCA	18	98.88	93.42
SVM-RFE	18.5	98.77	93.35

Table 13. Deep learning classification using SVs only

# of Features	Sensitivity	Specificity	Accuracy
Full	0	99.78	93.05
Chi Square	0	99.6	92.88
Gini Index	0.5	99.86	93.15
PCA	0	100	93.25
SVM-RFE	0	99.57	92.84

Table 14. Deep learning classification *SVs(P)*

# of Features	Sensitivity	Specificity	Accuracy
Full	55	81.79	79.98
Chi Square	90	30.85	34.84
Gini Index	69.5	83.16	82.24
PCA	41.5	91.78	88.39
SVM-RFE	81.5	73.9	74.41

Further, classification prediction experimentations are carried out using various training sets and compared with the classification results obtained using SVs(P). Table 6 to 8 shows the results yielded using Naïve Bayes classifier as a predictive model using full training data, SVs and *SVs(P),* including full features and reduced features.

Table 7 and Table 8 indicate that when extracted SVs are considered for training, Naïve Bayes performs best using PCA-based reduced feature data. Further, it yielded the best sensitivity of 96% using SVM-RFE-based reduced feature data with *SVs(P)*. Except for the PCA-based reduced features, Naïve Bayes tends to achieve similar accuracy when full feature data is used or reduced feature data is used. Tables 9 to 11 present the sensitivity yielded using the DT classifier with full data, SVs and *SVs(P),* including full features and reduced features. Results in Table 9 show that unlike Naïve Bayes performance using original training data, Decision Tree failed to perform well with original training data tested. Instead of using *SVs(P)* data, Decision Tree performance seems considerably improved.

It is observed that decision trees yielded similar and bad results using original training and extracted SVs (using full-featured data or reduced-featured data). It is observed from the results presented in Table 11 that reduced features using SVM-RFE or Gini Index and *SVs(P)* yielded the best sensitivity of 74.5% and 74%, respectively.

Tables 12 to 14 present the sensitivities yielded using Deep Learning classification using full training data, SVs and *SVs(P)* data, including full features and reduced features, respectively. It is observed that Deep Learning classifiers outperform with the highest sensitivity of 81.5% when trained with proposed modified data and SVM-RFE-based reduced features.

It should also be noted from the yielded results that Deep Learning classifiers perform better when full training data is used when compared to its performance with only SVs data.

Tables 15 to 17 present the sensitivity yielded using the Logistic Regression classifier with full data, SVs and *SVs(P)*, including full features and reduced features, respectively. It is observed from the results yielded that logistic regression failed to perform any better when full training data and extracted SVs were used for training. Whereas, using *SVs(P)*, it is observed that logistic regression outperformed with 91% sensitivity in this category with SVM-RFE-based reduced features.

Figures 2 to 5 represent the only sensitivities obtained by Naïve Bayes, Decision Tree, Deep Learning and Logistic Regression using full feature data and reduced feature data as well as original training data, SVs and SVs(P), respectively. Except for the Naïve Bayes classifier, all other classifiers tested in this study tend to perform best using the proposed modified data, i.e., SVs(P) for training.

Further, it is also observed that using the proposed data set SVs(P) with SVM-RFE-based reduced features yielded the best sensitivity of 96%. The results yielded using proposed modified data, i.e., SVs(P), are better compared to other training data sets tested in this study.

Table 15. Logistic regression classification full training data

# of Features	Sensitivity	Specificity	Accuracy
Full	2.5	99.49	92.94
Chi Square	0	99.95	93.21
Gini Index	1.5	99.82	93.18
PCA	0.5	99.89	93.18
SVM-RFE	3	99.64	93.11

Table 16. Logistic regression classification using SVs only

# of Features	Sensitivity	Specificity	Accuracy
Full	0.5	99.17	92.51
Chi Square	0	99.93	93.18
Gini Index	0.5	99.67	92.98
PCA	1	99.53	92.88
SVM-RFE	0	100	93.25

Table 17. Logistic regression classification using *SVs(P)*

# of Features	Sensitivity	Specificity	Accuracy
Full	66	82.84	81.7
Chi Square	13	97.18	91.49
Gini Index	66.5	80.74	79.78
PCA	17.5	86.35	81.7
SVM-RFE	**91**	25.13	29.57

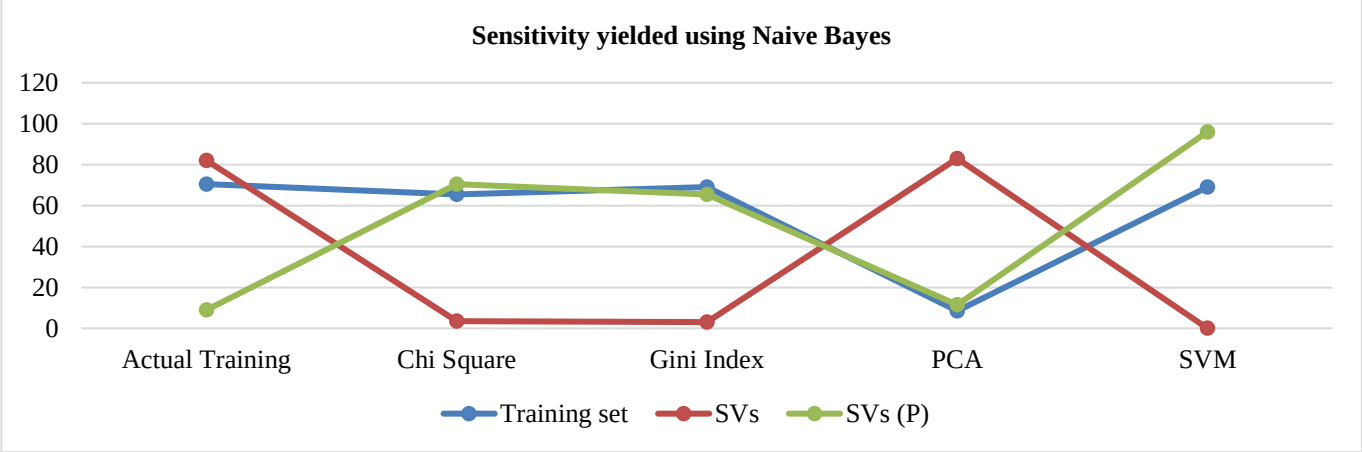

Fig. 2 Sensitivity yielded using Naïve Bayes classifier on the validation set

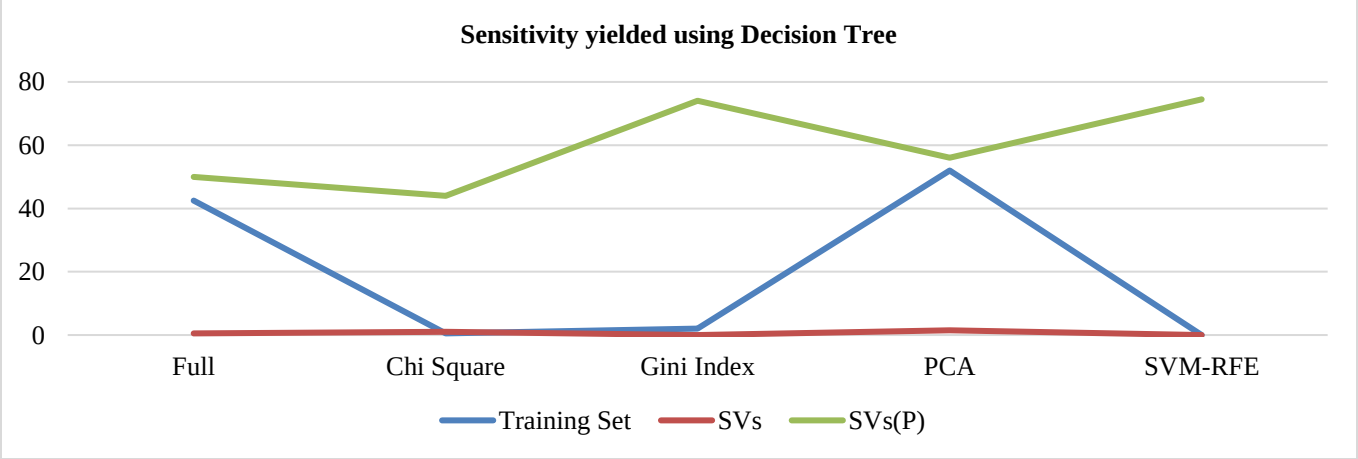

Fig. 3 Sensitivity yielded using decision tree classifier on the validation set

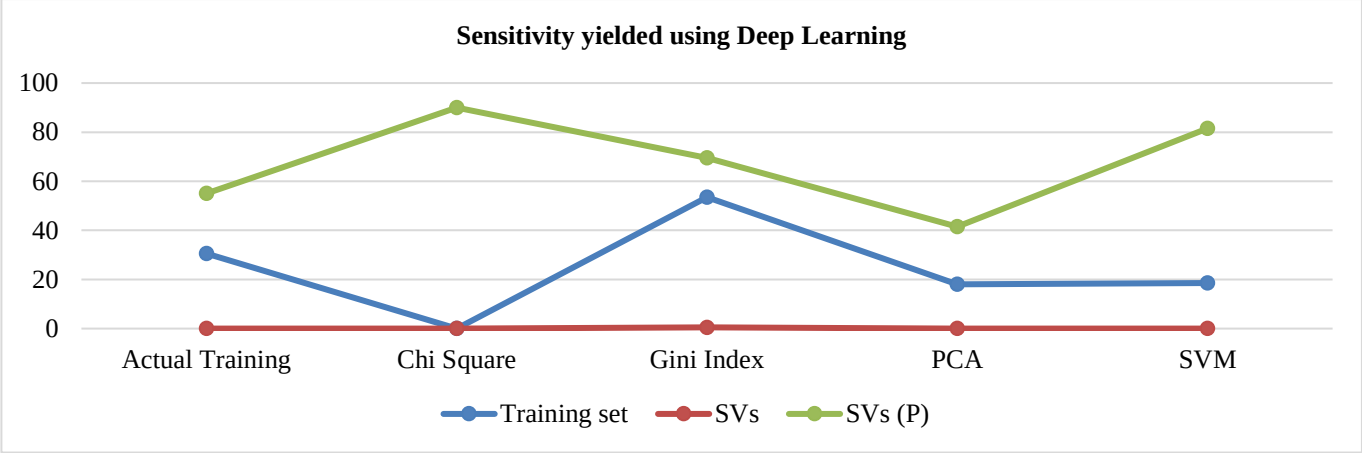

Fig. 4 Sensitivity yielded using deep learning classifier on the validation set

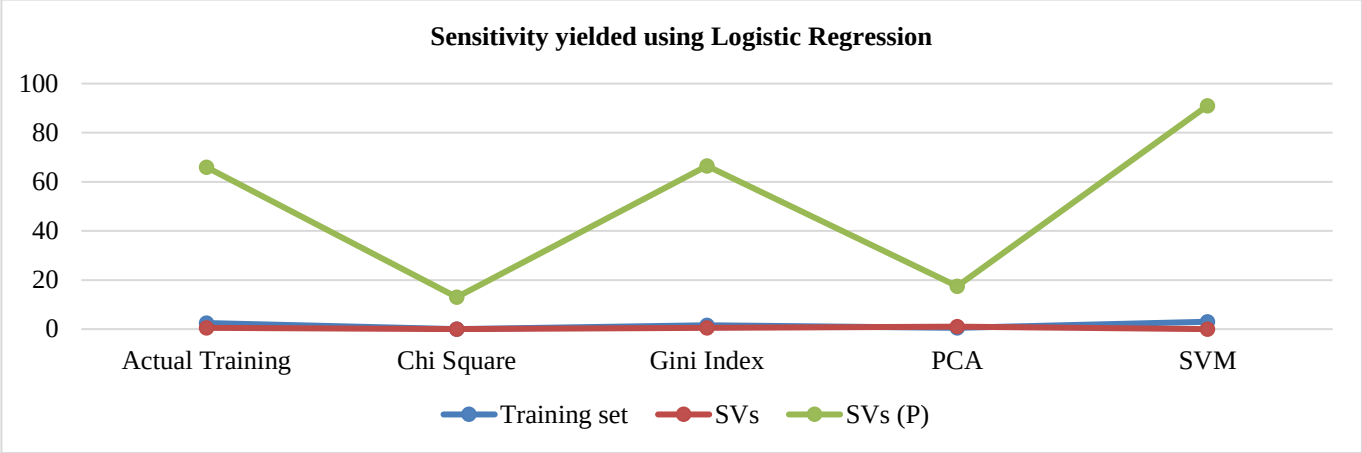

Fig. 5 Sensitivity yielded using logistic regression classifier on validation set

7. Conclusion

Soft Computing algorithms tend to give suboptimal models when training data is imbalanced. Using imbalanced data, learning of the algorithms is biased towards majority class instances. This paper proposes a data modification approach using SVM to deal with data imbalance problems. The problem analyzed in this study pertains to churn prediction in bank credit cards with a 93:6 class distribution ratio. The proposed approach extracts SVs using SVM. Later, target values of the extracted SVs are replaced by the predictions of trained SVM to obtain modified *SVs(P)* data.

Further, we have employed feature ranking using Chi-Square statistics, Gini Index, PCA and SVM-RFE to simplify the model-building process. Sensitivity is accorded the highest priority for all the experiments in this study. It is observed that soft computing approaches viz., Decision Tree, Deep Learning and Logistic Regression outperform when proposed modified data, i.e., *SVs(P)* is used. As an exception, Naïve Bayes tends to perform better with extracted original SVs when full feature data is used; nevertheless, it performs best when *SVs(P)* with reduced feature data are used.

Furthermore, it is observed that class distribution in *SVs(P)* sets is more balanced when compared to the original data and extracted SVs. Therefore, *SVs(P)* data helped soft computing algorithms learn better about minority class instances and yielded better results. In future, extensive research studies can be carried out using other soft computing techniques and a k-fold cross-validation approach using other datasets.

References

[1] Haibo He, and Edwardo A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[2] Liuzhi Yin et al., "Feature Selection for High-Dimensional Imbalanced Data," *Neurocomputing*, vol. 105, pp. 3-11, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[3] N.V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[4] R. Barendela et al., "Strategies for Learning in Class Imbalance Problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849-851, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[5] Gustavo E.A.P.A. Batista, Maria C. Monard, and Ana L.C. Bazzan, "Improving Rule Induction Precision for Automated Annotation by Balancing Skewed Data Sets," *Knowledge Exploration in Life Science Informatics (KELSI)*, pp. 20-32, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[6] Hongyn Guo, and Herna L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation: The Data Boosting Approach," *ACM SIGKDD Explorations*, vol. 6, no. 1, pp. 30-39, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[7] Taeho Jo, and Nathalie Japkowicz, "Class Imbalances Versus Small Disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40-49, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[8] Gilles Cohen et al., "Learning from Imbalanced Data in Surveillance of Nosocomial Infection," *Artificial Intelligence in Medicine*, vol. 37, no. 1, pp. 7-18, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[9] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *International Conference on Intelligent Computing (ICIC 2005), Advances in Intelligent Computing*, pp. 878-887, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yang Liu et al., "A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection In Speech," *Computer Speech and Language*, vol. 20, no. 4, pp. 468-494, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[11] Alberto Fernández, María José Del Jesus, and Francisco Herrera, "On the 2-Tuples Based Genetic Tuning Performance for Fuzzy Rule Based Classification Systems in Imbalanced Datasets," *Information Sciences*, vol. 180, no. 8, pp. 1268-1291, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[12] Yang Liu et al., "Combining Integrated Sampling with SVM Ensembles for Learning from Imbalanced Datasets," *Information Processing and Management*, vol. 47, no. 4, pp. 617-631, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mina Alibeigi, Sattar Hashemi, Ali Hamzeh, "DBFS: An Effective Density Based Feature Selection Scheme for Small Sample Size and High Dimensional Imbalanced Data Sets," *Data and Knowledge Engineering*, vol. 81, pp. 67-103, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ming Gao et al., "A Combined SMOTE and PSO Based RBF Classifier for Two-Class Imbalanced Problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456-3466, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[15] Adnan Idris, Muhammad Rizwan, and Asifullah Kham, "Churn Prediction in Telecom Using RANDOM Forest and PSO Based Data Balancing in Combination with Various Feature Selection Strategies," *Computers and Electrical Engineering*, vol. 38, no. 6, pp. 1808-1819, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[16] Sukarna Barua et al., "MWMOTE-Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405-425, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[17] Chris Seiffert et al., "An Empirical Study of the Classification Performance of Learners on Imbalanced and Noisy Software Quality Data," *Information Sciences*, vol. 259, pp. 571-595, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[18] Annarita D'Addabbo, and Rosalia Maglietta, "Parallel Selective Sampling Method for Imbalanced and Large Data Classification," *Pattern Recognition Letters*, vol. 62, pp. 61-67, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[19] Shounak Datta, and Swagatam Das, "Near-Bayesian Support Vector Machines for Imbalanced Data Classification with Equal or Unequal Misclassification Costs," *Neural Networks*, vol. 70, pp. 39-52, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[20] Zhongbin Sun et al., "A Novel Ensemble Method for Classifying Imbalanced Data," *Pattern Recognition*, vol. 48, no. 5, pp. 1623-1637, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[21] Min Zhu et al., "Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data," *IEEE Access*, vol. 6, pp. 4641-4652, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Chih-Fong Tsai et al., "Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection," *Information Sciences*, vol. 477, pp. 47-54, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Min Li et al., "ACO Resampling: Enhancing the Performance of Oversampling Methods for Class Imbalance Classification," *Knowledge Based Systems*, vol. 196, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] M. Aldiki Febriantono et al., "Classification of Multiclass Imbalanced Data Using Cost-Sensitive Decision Tree C50," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 65-72, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Mylam Chinnappan Babu, and Sangaralingam Pushpa, "Genetic Algorithm-Based PCA Classification for Imbalanced Dataset," *Intelligent Computing in Engineering*, pp. 541-552, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Jong Hyok Ri, and Hun Kim, "G-mean Based Extreme Learning Machine for Imbalance Learning," *Digital Signal Process*, vol. 98, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[27] Seba Susan, and Amitesh Kumar, "Hybrid of Intelligent Minority Oversampling and PSO-based Intelligent Majority Undersampling for Learning From Imbalanced Datasets," *International Conference on Intelligent Systems Design and Applications, ISDA 2018*, pp. 760-769, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[28] Arwa A. Jamjoom, "The Use of Knowledge Extraction in Predicting Customer Churn in B2B," *Journal of Big Data*, vol. 8, no. 110, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Praveen Lalwani et al., "Customer Churn Prediction System: A Machine Learning Approach," *Computing*, vol. 104, no. 8, pp. 271-294, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[30] Takuma Kimura, "Customer Churn Prediction with Hybrid Resampling and Ensemble Learning," *Journal of Management Information and Decision Sciences*, vol. 25, no. 1, pp. 1-23, 2022. [Google Scholar] [Publisher Link]

[31] Rencheng Liu et al., "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms," *Applied Sciences*, vol. 12, no. 18, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[32] Yan Zhang, and Lin Chen, "A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-sampling Method," *Theoretical Economics Letters*, vol. 11, no. 2, pp. 258-267, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[33] Business Intelligence Cup - 2004: Organized by the University of Chile. [Online]. Available: http://www.tis.cl/bicup_04/text-bicup/BICUP/202004/20public/20data.zip.

[34] Mierswa, and R. Klinkenberg, RapidMiner Studio, RapidMiner Account, 2018. [Online]. Available : https://rapidminer.com/

[35] V.P. Eswaramurthy, and S. Induja, "A Study on Customer Rentention Using Predictive Data Mining Techniques," *International Journal of Computer and organization Trends (IJCOT),* vol. 4, no. 5, pp. 6-10, 2014. [CrossRef] [Publisher Link]