International Journal of Sciences:
Basic and Applied Research
(IJSBAR)

ISSN 2307-4531
(Print & Online)

https://gssrr.org/index.php/JournalOfBasicAndApplied/index

International Journal of
Sciences:
Basic and Applied
Research

ISSN 2307-4531
(Print & Online)

Published by:

Evaluation of the Factors Affecting Classification Performance in Class Imbalance Problem

Duygu Aydin Hakli[a*], Dincer Goksuluk[b], Erdem Karabulut[c]

[a]Assist. Prof., Istanbul Arel University, Faculty of Medicine, Department of Biostatistics, Postcode 34010, Istanbul, Türkiye

[b]Assist. Prof., Erciyes University, Faculty of Medicine, Department of Biostatistics, Postcode 38030, Kayseri, Türkiye

[c]Prof. , Erciyes University, Faculty of Medicine, Department of Biostatistics, Postcode 06230, Ankara, Türkiye

[a]Email: duyguaydinhakli@gmail.com, [b]Email: dincergoksuluk@erciyes.edu.tr

[c]Email: ekarabul@hacettepe.edu.tr

Abstract

In binary classification, when the distribution of numbers in the class is imbalanced, we are aimed to increase the accuracy of classification in classification methods. In our study, simulated data sets and actual data sets are used. In the simulation, the "BinNor" package in the R project, which produces both numerical and categorical data, was utilized. When simulation work is planned, three different effects are considered which may affect the classification performance. These are: sample size, correlation structure and class imbalance rates. Scenarios were created by considering these effects. Each scenario was repeated 1000 times and 10-fold cross-validation was applied. CART, SVM and RF methods have been used in the classification of data sets obtained from both simulation and actual data sets. SMOTE, SMOTEBoost and RUSBoost were used to decrease or completely remove the imbalance of the data before the classification methods were applied. Specificity, sensitivity, balanced accuracy and F-measure were used as performance measures. The simulation results: the imbalance rate increases from 10 to 30, the effect of the 3 algorithms on the classification methods is similar accuracy. Because the class imbalance has become balanced.

Keywords: classification; imbalance learning; machine learning; oversampling; undersampling.

Received: 8/1/2023
Accepted: 9/5/2023
Published: 9/17/2023

* Corresponding author.

1. Introduction

Analyzing imbalanced data sets via machine learning algorithms has become a common and remarkable research area in recent years. Although it is possible to work with balanced data sets in simulation studies, it is more likely to be work with imbalanced data sets in real life examples. In a binary classification, for example, class imbalance problem emerges when one of the classes has few instances compared to the other class [1, 8]. This problem is very common in Health Sciences, especially in rare event studies, cancer studies and etc. As an example, the level of imbalance in a case-control study (ratio of the size of control group to case group) might be 5%, 10% or 15%. The imbalance problem heightens whenever the class of relevant has a small number of instances compared to the majority class [1, 3]. While the class with a higher set of instances is called as a “majority class”, the other class is called as a “minority class” [1]. We use the terms “minority” and “majority” throughout this paper.

The imbalanced data sets lead to several problems such as lower model accuracies, convergence problem for some of the learning algorithms and overfitting (or underfitting) [4, 9, 12]. The subjects in minority class are misclassified since the model is trained in favor of majority class. As a result, accuracy of the model decreases due to the misclassified instances in minority class [1, 4, 6, 8]. It is possible to use balanced accuracy as the measure of model performance; however, the results are still sensitive to the unequal sample sizes when the level of imbalance is very low (e.g, 5 to 10 percent). When the imbalanced dataset is directly used in classification task, the accuracy of the classification methods might be decreased. There are several approaches in the literature which might be preferred to decrease or eliminate the effect of the class imbalances. Some of these solutions are (i) creating synthetic data, (ii) undersampling and (iii) oversampling [1]. In the former approach, new samples are generated using the information from data at hand. In the latter approaches, the basic idea is to use resampled subjects from data [1, 5]. Undersampling aims to equalize the class sizes by randomly removing instance from majority group [4, 6, 7, 13]. Oversampling, on the other hand, focuses on the minority group and aims to equalize the class sizes by replicating subject in the corresponding class [1, 4, 6, 7, 13]. Several algorithms based on three ideas described above are proposed in order to overcome class imbalance problem. Among these algorithms, SMOTE (Synthetic Minority Oversampling Technique) [1], SMOTEBoost (SMOTE with Boosting) [5] and RUSBoost (Random undersampling with Boosting) [6] were considered in this study.

Chawla and his colleagues [1] proposed a synthetic minority oversampling technique (SMOTE) by generating synthetic instances to handle this problem. Another algorithm, SMOTEBoost [5], aims solve the class imbalance problem by combining boosting algorithms with SMOTE. Instead of changing the distributions of training data by updating the weights associated with each instance, SMOTEBoost alters the distributions by adding new instances of minority class using the SMOTE algorithm; however, it is time-consuming sampling technique. Seiffert and his colleagues [6] proposed a new algorithm, RUSBoost, which combines RUS and boosting procedure. Both RUSBoost and SMOTEBoost are basically based on boosting algorithm. RUSBoost method requires shorter time than SMOTEBoost for training the model. For this reason, this method is preferred over SMOTE and SMOTEBoost.

Three data balancing algorithms were used and then the classification accuracies are compared with each other. These classification methods are Support Vector Machines (SVM), Classification and Regression Tree (CART) and Random Forest (RF) [14, 17]. SVM is one of the well-known methods in the literature to separate classes that cannot be separated linearly [8]. Classification is an important part of machine learning. In the classification task, data is divided into two parts: 70% in the training set and 30% in the test set. The model is fitted to training set and optimum model parameters are obtained using validation techniques such as cross-validation, bootstrap, etc. Then, test set is used to measure the test performances of the trained models [14].

In this article, we focus on binary classification study under several conditions which are likely to affect the model accuracies when data are unbalanced. In the previous studies, different sample sizes and different correlation structures (the relationship between independent variables and dependent variable) were not considered. The level of the correlation between dependent and independent variables may also affect the classification results. Therefore, we compared the results of these algorithms on different imbalance ratios, sample sizes and correlation structures using a simulation study. The results revealed that accuracies of classification methods are increased while using with proposed data balancing algorithms.

This paper organized as follows: Section 2 the algorithms contain that used to solve the problem of imbalance dataset. Section 3 considers performance measures. Section 4 gives details of a simulation study and real datasets. Section 5 provides results of experiments. Finally, we conclude in Section 6.

2. Data Balancing Algorithms

2.1. SMOTE (Synthetic Minority Oversampling Technique)

SMOTE [1] technique is an oversampling method which is based on the idea that generating the synthetic minority examples. The algorithm in the background can be briefly described as below:

Step 1: search for the k nearest neighbors of each minority instance,

Step 2: take the difference between original minority class instance (x_i) and the k nearest neighborhood (kNN) instance (x_j),

Step 3: select a random number (α) between [0, 1] and multiply the difference obtained from step 2 by α,

Step 4: create a new synthetic instance using the equation (1).

$$x_{new}=x_i + (x_j-x_i) * \alpha \qquad (1)$$

Step 5. Repeat steps 1-4 iteratively to generate desired number of synthetic instances.

2.2. SMOTEBoost (SMOTE + Boosting)

SMOTEBoost [5] algorithm combines SMOTE and standard boosting (resampling method) procedures.

This method has two major advantages over SMOTE. First, the standard boosting procedure gives equal weights to all misclassified instances.

In SMOTEBoost algorithm, weights are changed indirectly because synthetic instances generated from the minority class. In addition, skewed distribution is balanced. Second, SMOTEBoost increases True Positive Rate (TPR), so the accuracy of the model is improved.

2.3. RUSBoost (Random Undersampling + Boosting)

RUS (Random Undersampling) [6] is a method which aims to equalize the number of samples in each class by selecting a random sample from majority group.

The remaining instances are removed from dataset; thus, the number of instances in each group is set to be equal (or almost equal). RUSBoost combines RUS and standard boosting procedures. The algorithm is as below:

Step 1. Let m be the number of total instances in the training set. Weight of each instance is set to be 1/m.

Step 2. For T times, a training set is created by applying RUS method on the original majority set.

3. Performance Measures

In a binary classification task, common metrics are derived from a confusion matrix as shown in Table 1. In this paper, we used Balanced Accuracy and F-Measure as performance measures since these measures take imbalanced number of samples into account.

Table 1: Confusion matrix for a binary classification.

		REAL		**TOTAL**
		Present	**Absent**	
PREDICTION	**Positive**	True Positive(TP)	False Positive(FP)	**TP+FP**
	Negative	False Negative(FN)	True Negative(TN)	**FN+TN**
TOTAL		TP+FN	FP+TN	**N**

$$\text{Sensitivity}= TP/(TP+FN) \tag{2}$$

$$\text{Specificity}= TN/(FP+TN) \tag{3}$$

$$\text{Precision}=TP/(TP+FP) \tag{4}$$

$$\text{Recall}=TP/(TP+FN) \tag{5}$$

$$\text{F-Measure}=2\ ((\text{Precision}*\text{Recall})/(\text{Precision} + \text{Recall})) \tag{6}$$

$$\text{Balanced Accuracy}=1/(2(\text{Sensitivity} + \text{Specificity})) \tag{7}$$

4. Experiments

4.1. Simulation Study

In the literature, many studies which are related to class imbalance problem are published. Some of these studies based on simulation study, however, these simulation studies were not examined under different correlation structures. Therefore; we believe that this study will contribute to demonstrate the validity of proposed algorithms under different correlation structures. The accuracy of classifiers might be affected by the sample size, correlation structures and class imbalance rates. Hence, simulation scenarios were based on these effects. The simulation study contains all possible combinations of:

1) correlation structures as low, medium, high and actual (i.e the correlation structure obtained from real dataset)
2) sample sizes as 100, 250, 500, 1000 and 2000
3) imbalance ratios as 10%, 15%, 25% and 30%

The correlation between dependent and independent variables for low, medium and high correlation structures are defined in the intervals [0.00, 0.20], [0.30, 0.60] and [0.61, 0.90] respectively. We assume that weak or no correlation exist among independent variables. Fifteen variables (5 binary, 10 numeric) along with the class variable are generated using "BinNor" [18] package from R program (version 3.2.3). All scenarios (i.e 80 different combinations) were repeated 1000 times and 5-fold cross-validation was applied to find optimum model parameters. Simulated and real datasets were trained using CART, SVM and RF methods with default options in R. SMOTE, SMOTEBoost and RUSBoost were used to decrease or completely remove the imbalance of the data before the classification methods were applied. In addition, we applied only classification methods without balancing algorithms. This is expressed as NONE in the tables. We used "caret" [19], "rpart" [20] and "ROSE" [21] packages from R program (version 3.2.3). Figure 1 shows the workflow of the simulation study.

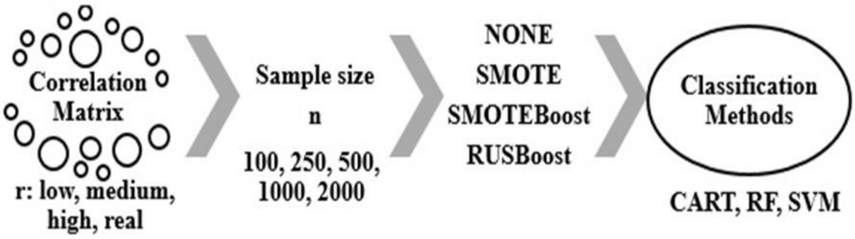

Figure 1: Summary of simulation study.

4.2. Real Datasets

Our experiments were performed on the seven real datasets summarized in Table 2. We received six datasets from open source website: http://archive.ics.uci.edu/ml/ (UC Irvine Machine Learning Repository). One of them received from the article deducted as a note at Table 2.

Table 2: Summary of Real Datasets used in experiments.

Datasets	Sample Size	Minority class/Total	Number of Variables
Abalone	731	%5,74- %94,26	9
Fertility	100	%12-%88	5
Thoracis Surgery	470	%15-%85	16
Hepatitis	155	%20-%80	19
Blood transfusion	748	%23-%77	5
Alzheirmer*	70	%30-%70	26
Pima Diabetes	768	%34,9-%66,1	9

* A blood based 12-miRNA signature of Alzheimer disease patients. (2013), Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, and his colleagues.

5. Results

5.1. Simulation Results

The results given in tables were evaluated according to F-measure and Balanced Accuracy.

Low-level correlation;

When the imbalance ratio was 10%, according to the F-measure values, SMOTE and SMOTEBoost algorithms were similar impacts on the classification accuracy. The balanced accuracy of SVM and RF were approximately 30% when RUSBoost was performed (Table 3).

As the imbalance rate increased to 30%, the effects of the 3 balancing algorithms on the classification methods were similar. This result is due to moderate imbalance rate. Hence, as the imbalance rate increases the change in model accuracies due to balancing algorithm decreases.

Table 3: Low-level correlation results.

CM	Minority/Sample size	Algorithms	Performance Measures									
			Balanced Accuracy					F-Measure				
			100	250	500	1000	2000	100	250	500	1000	2000
SVM	0,1	NONE	0,505	0,502	0,501	0,501	0,099	0,173	0,104	0,079	0,049	0,035
	0,1	RUSBoost	0,572	0,595	0,63	0,643	0,657	0,192	0,222	0,266	0,354	0,362
	0,1	SMOTE	0,506	0,506	0,503	0,508	0,503	0,143	0,116	0,066	0,065	0,035
	0,1	SMOTEBoost	0,506	0,505	0,503	0,508	0,503	0,129	0,088	0,06	0,064	0,034
	0,15	NONE	0,503	0,504	0,502	0,502	0,502	0,14	0,131	0,088	0,054	0,043
	0,15	RUSBoost	0,559	0,601	0,602	0,622	0,631	0,179	0,284	0,323	0,367	0,449
	0,15	SMOTE	0,52	0,517	0,516	0,512	0,52	0,167	0,129	0,105	0,077	0,097
	0,15	SMOTEBoost	0,519	0,516	0,516	0,512	0,52	0,153	0,12	0,103	0,076	0,097
	0,25	NONE	0,51	0,507	0,506	0,506	0,25	0,228	0,171	0,132	0,093	0,073
	0,25	RUSBoost	0,556	0,572	0,595	0,6	0,605	0,376	0,439	0,449	0,481	0,478
	0,25	SMOTE	0,535	0,55	0,556	0,561	0,562	0,263	0,298	0,306	0,314	0,314
	0,25	SMOTEBoost	0,536	0,551	0,556	0,561	0,562	0,255	0,298	0,305	0,314	0,314
	0,3	NONE	0,51	0,511	0,512	0,512	0,513	0,245	0,205	0,163	0,134	0,117
	0,3	RUSBoost	0,561	0,57	0,589	0,597	0,601	0,392	0,463	0,463	0,5	0,492
	0,3	SMOTE	0,542	0,56	0,567	0,575	0,58	0,355	0,395	0,393	0,406	0,417
	0,3	SMOTEBoost	0,543	0,558	0,567	0,574	0,58	0,358	0,395	0,393	0,406	0,418
CART	0,1	NONE	0,5	0,502	0,503	0,503	0,099	0,128	0,096	0,098	0,082	0,058
	0,1	RUSBoost	0,5	0,555	0,557	0,566	0,577	0	0,162	0,228	0,286	0,299
	0,1	SMOTE	0,51	0,506	0,503	0,506	0,503	0,143	0,115	0,099	0,101	0,074
	0,1	SMOTEBoost	0,508	0,505	0,501	0,502	0,501	0,099	0,017	0,014	0,015	0,007
	0,15	NONE	0,502	0,505	0,505	0,505	0,503	0,19	0,147	0,116	0,095	0,079
	0,15	RUSBoost	0,529	0,541	0,545	0,558	0,565	0,171	0,231	0,322	0,312	0,362
	0,15	SMOTE	0,517	0,513	0,513	0,509	0,519	0,195	0,167	0,158	0,13	0,128
	0,15	SMOTEBoost	0,513	0,506	0,507	0,503	0,515	0,095	0,024	0,059	0,026	0,098
	0,25	NONE	0,511	0,507	0,507	0,505	0,25	0,247	0,196	0,177	0,147	0,133
	0,25	RUSBoost	0,529	0,532	0,545	0,547	0,552	0,343	0,382	0,39	0,417	0,422
	0,25	SMOTE	0,524	0,529	0,533	0,535	0,537	0,284	0,288	0,281	0,274	0,266
	0,25	SMOTEBoost	0,526	0,525	0,529	0,532	0,533	0,245	0,211	0,218	0,227	0,22
	0,3	NONE	0,508	0,513	0,513	0,51	0,507	0,273	0,257	0,219	0,19	0,171
	0,3	RUSBoost	0,53	0,532	0,543	0,544	0,548	0,305	0,423	0,398	0,45	0,443
	0,3	SMOTE	0,527	0,537	0,542	0,544	0,548	0,355	0,368	0,356	0,354	0,361
	0,3	SMOTEBoost	0,527	0,538	0,544	0,545	0,549	0,331	0,359	0,346	0,34	0,365
RF	0,1	NONE	0,505	0,506	0,503	0,504	0,099	0,179	0,113	0,059	0,046	0,03
	0,1	RUSBoost	0,562	0,59	0,613	0,627	0,642	0,151	0,169	0,347	0,335	0,353
	0,1	SMOTE	0,508	0,504	0,508	0,515	0,511	0,118	0,071	0,067	0,078	0,057
	0,1	SMOTEBoost	0,506	0,505	0,506	0,512	0,508	0,067	0,042	0,037	0,056	0,038
	0,15	NONE	0,508	0,508	0,507	0,506	0,505	0,154	0,095	0,066	0,043	0,033
	0,15	RUSBoost	0,556	0,588	0,582	0,603	0,615	0,231	0,31	0,325	0,368	0,434
	0,15	SMOTE	0,524	0,524	0,52	0,52	0,527	0,168	0,131	0,114	0,107	0,131
	0,15	SMOTEBoost	0,519	0,522	0,518	0,516	0,523	0,12	0,106	0,093	0,084	0,11
	0,25	NONE	0,52	0,521	0,521	0,518	0,25	0,174	0,14	0,125	0,106	0,096
	0,25	RUSBoost	0,555	0,56	0,578	0,585	0,59	0,356	0,412	0,44	0,463	0,467
	0,25	SMOTE	0,534	0,547	0,553	0,557	0,562	0,261	0,294	0,297	0,307	0,314
	0,25	SMOTEBoost	0,534	0,55	0,553	0,558	0,561	0,248	0,287	0,291	0,302	0,309
	0,3	NONE	0,523	0,526	0,529	0,53	0,529	0,201	0,188	0,189	0,182	0,171
	0,3	RUSBoost	0,558	0,559	0,572	0,581	0,587	0,411	0,441	0,458	0,485	0,483
	0,3	SMOTE	0,543	0,553	0,565	0,571	0,577	0,357	0,383	0,388	0,4	0,413
	0,3	SMOTEBoost	0,541	0,554	0,565	0,572	0,578	0,348	0,379	0,384	0,399	0,412

CM:Classification Methods

Medium-level correlation;

When the imbalance ratio was 10%, the RUSBoost algorithm increased the classification accuracies. Especially the classification accuracy of SVM and RF were increased up to 70% with the increasing size of samples (Table 4).

SMOTE and SMOTEBoost algorithms had also increased the model accuracies, however, RUSBoost's effect was greater.

Table 4: Medium-level correlation results.

CM	**Minority/Sample size**	**Algorithms**	**Performance Measures**									
			Balanced Accuracy					**F-Measure**				
			100	250	500	1000	2000	100	250	500	1000	2000
SVM	0,1	NONE	0,562	0,612	0,643	0,099	0,694	0,4	0,4	0,415	0,465	0,506
	0,1	RUSBoost	0,752	0,809	0,826	0,847	0,863	0,531	0,632	0,675	0,712	0,739
	0,1	SMOTE	0,598	0,649	0,641	0,705	0,739	0,408	0,398	0,384	0,5	0,558
	0,1	SMOTEBoost	0,599	0,651	0,642	0,705	0,739	0,396	0,397	0,384	0,5	0,558
	0,15	NONE	0,58	0,617	0,148	0,676	0,15	0,399	0,407	0,508	0,48	0,584
	0,15	RUSBoost	0,725	0,78	0,796	0,818	0,829	0,545	0,675	0,697	0,699	0,725
	0,15	SMOTE	0,677	0,73	0,753	0,766	0,768	0,458	0,546	0,58	0,595	0,608
	0,15	SMOTEBoost	0,68	0,731	0,754	0,767	0,768	0,458	0,548	0,581	0,596	0,608
	0,25	NONE	0,619	0,247	0,688	0,249	0,712	0,449	0,53	0,532	0,593	0,579
	0,25	RUSBoost	0,717	0,755	0,78	0,788	0,808	0,659	0,684	0,704	0,71	0,704
	0,25	SMOTE	0,726	0,752	0,778	0,79	0,803	0,577	0,619	0,655	0,668	0,682
	0,25	SMOTEBoost	0,725	0,752	0,778	0,79	0,803	0,575	0,62	0,655	0,668	0,682
	0,3	NONE	0,295	0,679	0,299	0,718	0,725	0,524	0,537	0,624	0,603	0,616
	0,3	RUSBoost	0,725	0,75	0,772	0,776	0,797	0,645	0,689	0,712	0,727	0,727
	0,3	SMOTE	0,726	0,754	0,778	0,792	0,802	0,608	0,65	0,684	0,701	0,716
	0,3	SMOTEBoost	0,728	0,754	0,779	0,791	0,802	0,611	0,65	0,684	0,701	0,716
CART	0,1	NONE	0,519	0,56	0,577	0,099	0,605	0,322	0,317	0,32	0,332	0,342
	0,1	RUSBoost	0,5	0,689	0,71	0,738	0,75	0	0,459	0,491	0,57	0,592
	0,1	SMOTE	0,571	0,61	0,584	0,624	0,643	0,311	0,324	0,307	0,355	0,383
	0,1	SMOTEBoost	0,571	0,595	0,563	0,621	0,644	0,247	0,253	0,19	0,326	0,379

	0,15	NONE	0,54	0,566	0,148	0,602	0,15	0,353	0,337	0,53	0,352	0,573
	0,15	RUSBoost	0,654	0,72	0,748	0,771	0,791	0,393	0,573	0,658	0,668	0,696
	0,15	SMOTE	0,654	0,714	0,745	0,764	0,767	0,414	0,516	0,565	0,59	0,601
	0,15	SMOTEBoost	0,651	0,711	0,741	0,761	0,765	0,374	0,501	0,556	0,585	0,599
	0,25	NONE	0,579	0,247	0,615	0,249	0,637	0,406	0,512	0,416	0,581	0,444
	0,25	RUSBoost	0,643	0,707	0,723	0,741	0,758	0,526	0,604	0,64	0,686	0,693
	0,25	SMOTE	0,683	0,712	0,73	0,742	0,753	0,495	0,555	0,585	0,605	0,616
	0,25	SMOTEBoost	0,676	0,71	0,729	0,742	0,754	0,46	0,555	0,588	0,604	0,616
	0,3	NONE	0,295	0,616	0,299	0,644	0,653	0,496	0,454	0,583	0,48	0,496
	0,3	RUSBoost	0,646	0,702	0,722	0,737	0,75	0,501	0,592	0,625	0,666	0,689
	0,3	SMOTE	0,68	0,714	0,735	0,753	0,762	0,539	0,594	0,622	0,646	0,658
	0,3	SMOTEBoost	0,676	0,712	0,733	0,752	0,762	0,558	0,591	0,619	0,646	0,658
RF	0,1	NONE	0,566	0,6	0,614	0,099	0,641	0,388	0,329	0,342	0,375	0,406
	0,1	RUSBoost	0,735	0,788	0,796	0,813	0,828	0,497	0,614	0,643	0,676	0,692
	0,1	SMOTE	0,59	0,641	0,625	0,671	0,693	0,374	0,382	0,351	0,445	0,484
	0,1	SMOTEBoost	0,591	0,631	0,617	0,664	0,686	0,324	0,357	0,331	0,432	0,473
	0,15	NONE	0,587	0,609	0,148	0,634	0,15	0,359	0,341	0,459	0,399	0,545
	0,15	RUSBoost	0,68	0,743	0,775	0,805	0,821	0,486	0,631	0,675	0,688	0,707
	0,15	SMOTE	0,637	0,707	0,732	0,754	0,762	0,392	0,504	0,548	0,578	0,594
	0,15	SMOTEBoost	0,631	0,698	0,727	0,751	0,756	0,361	0,485	0,54	0,575	0,584
	0,25	NONE	0,627	0,247	0,654	0,249	0,678	0,411	0,471	0,47	0,581	0,519
	0,25	RUSBoost	0,671	0,725	0,757	0,775	0,794	0,592	0,651	0,684	0,699	0,702
	0,25	SMOTE	0,677	0,724	0,752	0,767	0,782	0,497	0,58	0,622	0,642	0,661
	0,25	SMOTEBoost	0,674	0,724	0,75	0,767	0,782	0,487	0,576	0,618	0,639	0,659
	0,3	NONE	0,295	0,658	0,299	0,68	0,691	0,423	0,497	0,593	0,542	0,56
	0,3	RUSBoost	0,671	0,721	0,751	0,764	0,783	0,595	0,652	0,687	0,704	0,715
	0,3	SMOTE	0,691	0,731	0,758	0,774	0,787	0,558	0,619	0,657	0,677	0,695
	0,3	SMOTEBoost	0,69	0,733	0,759	0,777	0,791	0,573	0,617	0,655	0,677	0,695

CM: Classification Methods

High-level correlation ;

The effect on the classification accuracy of the RUSBoost algorithm (nearly 80%) was higher than SMOTE and SMOTEBoost when the imbalance ratio was 10% (Table 5).

As the imbalance ratio was decreased, the effects of the classification methods of the 3 algorithms on the classification accuracies were becoming similar.

Table 5: High-level correlation results.

CM	Minority/Sample size	Algorithms	Performance Measures									
			Balanced Accuracy					F-Measure				
			100	250	500	1000	2000	100	250	500	1000	2000
SVM	0,1	NONE	0,091	0,095	0,511	0,792	0,1	0,528	0,574	0,163	0,692	0,699
	0,1	RUSBoost	0,727	0,795	0,806	0,809	0,807	0,628	0,748	0,803	0,815	0,828
	0,1	SMOTE	0,653	0,766	0,794	0,796	0,794	0,527	0,609	0,682	0,691	0,694
	0,1	SMOTEBoost	0,652	0,767	0,794	0,796	0,794	0,51	0,609	0,682	0,691	0,694
	0,15	NONE	0,643	0,709	0,721	0,724	0,72	0,493	0,553	0,597	0,611	0,607
	0,15	RUSBoost	0,722	0,787	0,8	0,811	0,825	0,603	0,71	0,743	0,8	0,84
	0,15	SMOTE	0,698	0,742	0,745	0,738	0,741	0,517	0,577	0,589	0,598	0,605
	0,15	SMOTEBoost	0,697	0,741	0,744	0,738	0,741	0,511	0,577	0,588	0,598	0,605
	0,25	NONE	0,241	0,722	0,249	0,747	0,25	0,552	0,585	0,612	0,632	0,64
	0,25	RUSBoost	0,754	0,818	0,84	0,849	0,851	0,696	0,708	0,717	0,722	0,714
	0,25	SMOTE	0,756	0,799	0,811	0,821	0,835	0,619	0,669	0,678	0,687	0,695
	0,25	SMOTEBoost	0,759	0,8	0,811	0,821	0,835	0,624	0,672	0,678	0,687	0,695
	0,3	NONE	0,295	0,747	0,299	0,299	0,759	0,603	0,644	0,657	0,663	0,666
	0,3	RUSBoost	0,764	0,82	0,836	0,845	0,3	0,691	0,74	0,745	0,752	0,749
	0,3	SMOTE	0,775	0,817	0,827	0,84	0,3	0,673	0,725	0,734	0,746	0,747
	0,3	SMOTEBoost	0,778	0,818	0,827	0,84	0,3	0,677	0,728	0,733	0,746	0,747
CART	0,1	NONE	0,091	0,095	0,514	0,797	0,1	0,5	0,67	0,219	0,699	0,702
	0,1	RUSBoost	0,5	0,79	0,806	0,808	0,81	0	0,718	0,784	0,806	0,819
	0,1	SMOTE	0,705	0,784	0,798	0,797	0,794	0,508	0,626	0,678	0,688	0,687
	0,1	SMOTEBoost	0,702	0,783	0,796	0,796	0,792	0,45	0,617	0,672	0,686	0,683
	0,15	NONE	0,61	0,71	0,72	0,72	0,72	0,55	0,58	0,60	0,61	0,60

			5	3	5	4	1	8	9	6	3	9
	0,15	RUSBoost	0,712	0,788	0,804	0,813	0,824	0,494	0,682	0,761	0,809	0,891
	0,15	SMOTE	0,72	0,751	0,747	0,738	0,738	0,527	0,577	0,592	0,597	0,613
	0,15	SMOTEBoost	0,699	0,753	0,749	0,74	0,74	0,459	0,582	0,599	0,601	0,62
	0,25	NONE	0,241	0,74	0,249	0,751	0,25	0,608	0,618	0,63	0,638	0,64
	0,25	RUSBoost	0,728	0,825	0,829	0,838	0,845	0,637	0,695	0,695	0,701	0,698
	0,25	SMOTE	0,777	0,82	0,827	0,832	0,849	0,637	0,682	0,685	0,692	0,701
	0,25	SMOTEBoost	0,771	0,815	0,825	0,832	0,848	0,624	0,676	0,684	0,692	0,7
	0,3	NONE	0,295	0,761	0,299	0,299	0,771	0,637	0,663	0,664	0,659	0,68
	0,3	RUSBoost	0,746	0,82	0,821	0,834	0,3	0,637	0,721	0,722	0,735	0,738
	0,3	SMOTE	0,789	0,829	0,837	0,843	0,3	0,685	0,733	0,741	0,748	0,748
	0,3	SMOTEBoost	0,782	0,82	0,834	0,842	0,3	0,673	0,72	0,738	0,747	0,747
RF	0,1	NONE	0,091	0,095	0,529	0,785	0,1	0,514	0,597	0,189	0,679	0,692
	0,1	RUSBoost	0,726	0,786	0,805	0,811	0,807	0,561	0,715	0,775	0,786	0,806
	0,1	SMOTE	0,674	0,767	0,787	0,793	0,792	0,507	0,597	0,654	0,666	0,68
	0,1	SMOTEBoost	0,663	0,765	0,786	0,792	0,792	0,443	0,589	0,652	0,664	0,677
	0,15	NONE	0,617	0,72	0,724	0,725	0,721	0,478	0,577	0,6	0,611	0,608
	0,15	RUSBoost	0,726	0,787	0,808	0,809	0,816	0,588	0,711	0,757	0,784	0,788
	0,15	SMOTE	0,686	0,738	0,741	0,738	0,743	0,49	0,566	0,584	0,592	0,598
	0,15	SMOTEBoost	0,675	0,736	0,737	0,735	0,739	0,444	0,563	0,574	0,583	0,585
	0,25	NONE	0,241	0,72	0,249	0,727	0,25	0,52	0,576	0,587	0,598	0,606
	0,25	RUSBoost	0,745	0,816	0,828	0,836	0,842	0,67	0,705	0,715	0,721	0,718
	0,25	SMOTE	0,747	0,788	0,79	0,798	0,813	0,601	0,658	0,662	0,671	0,683
	0,25	SMOTEBoost	0,743	0,791	0,79	0,798	0,814	0,591	0,661	0,661	0,67	0,683
	0,3	NONE	0,295	0,756	0,299	0,299	0,77	0,576	0,654	0,671	0,676	0,68
	0,3	RUSBoost	0,75	0,809	0,821	0,834	0,3	0,682	0,733	0,738	0,746	0,747
	0,3	SMOTE	0,768	0,805	0,809	0,826	0,3	0,663	0,712	0,718	0,734	0,737
	0,3	SMOTEBoost	0,768	0,806	0,809	0,826	0,3	0,661	0,712	0,717	0,735	0,738

CM: Classification Methods

<u>*Real dataset correlation;*</u>

Similar results were found with the results of the level of medium correlation (Table 6). RUSBoost's effect was greater than SMOTE and SMOTEBoost algorithms.

Table 6: Real dataset correlation results.

CM	**Minority/Sample size**	**Algorithms**	**Performance Measures**									
			Balanced Accuracy					**F-Measure**				
			100	250	500	1000	2000	100	250	500	1000	2000
SVM	0,1	NONE	0,578	0,646	0,676	0,099	0,706	0,458	0,449	0,474	0,509	0,537
	0,1	RUSBoost	0,774	0,795	0,827	0,841	0,854	0,558	0,677	0,702	0,704	0,733
	0,1	SMOTE	0,644	0,736	0,731	0,757	0,757	0,447	0,528	0,533	0,569	0,583
	0,1	SMOTEBoost	0,649	0,739	0,731	0,757	0,757	0,448	0,532	0,533	0,569	0,583
	0,15	NONE	0,138	0,146	0,148	0,149	0,704	0,428	0,452	0,485	0,519	0,544
	0,15	RUSBoost	0,747	0,794	0,779	0,812	0,819	0,588	0,674	0,676	0,707	0,717
	0,15	SMOTE	0,697	0,719	0,717	0,753	0,76	0,505	0,517	0,532	0,588	0,598
	0,15	SMOTEBoost	0,697	0,718	0,715	0,753	0,76	0,502	0,517	0,53	0,589	0,599
	0,25	NONE	0,642	0,247	0,691	0,249	0,711	0,48	0,505	0,539	0,568	0,58
	0,25	RUSBoost	0,733	0,745	0,759	0,772	0,783	0,659	0,653	0,669	0,7	0,704
	0,25	SMOTE	0,715	0,727	0,747	0,746	0,764	0,559	0,583	0,613	0,617	0,64
	0,25	SMOTEBoost	0,716	0,726	0,747	0,746	0,764	0,559	0,582	0,613	0,617	0,64
	0,3	NONE	0,646	0,685	0,701	0,714	0,3	0,496	0,542	0,574	0,597	0,611
	0,3	RUSBoost	0,72	0,741	0,753	0,762	0,773	0,622	0,663	0,666	0,69	0,698
	0,3	SMOTE	0,706	0,73	0,749	0,752	0,765	0,575	0,617	0,643	0,649	0,665
	0,3	SMOTEBoost	0,702	0,73	0,75	0,752	0,765	0,577	0,618	0,644	0,649	0,665
CART	0,1	NONE	0,511	0,578	0,59	0,099	0,644	0,383	0,348	0,357	0,382	0,418
	0,1	RUSBoost	0,5	0,749	0,739	0,763	0,777	0	0,552	0,543	0,557	0,618
	0,1	SMOTE	0,62	0,679	0,689	0,704	0,695	0,395	0,418	0,434	0,452	0,468
	0,1	SMOTEBoost	0,645	0,683	0,675	0,703	0,696	0,341	0,403	0,392	0,45	0,468
	0,15	NONE	0,138	0,146	0,148	0,149	0,655	0,39	0,382	0,391	0,414	0,445
	0,15	RUSBoost	0,615	0,666	0,676	0,732	0,739	0,33	0,493	0,528	0,588	0,605
	0,15	SMOTE	0,64	0,67	0,68	0,69	0,69	0,4	0,43	0,45	0,48	0,48

			5	3	4	9	5		7	2	6	4
	0,15	SMOTEBoost	0,633	0,662	0,672	0,7	0,695	0,352	0,372	0,423	0,488	0,484
	0,25	NONE	0,585	0,247	0,651	0,249	0,681	0,423	0,451	0,47	0,495	0,523
	0,25	RUSBoost	0,61	0,652	0,684	0,699	0,71	0,511	0,536	0,566	0,619	0,635
	0,25	SMOTE	0,649	0,669	0,689	0,7	0,717	0,468	0,502	0,528	0,546	0,571
	0,25	SMOTEBoost	0,649	0,675	0,69	0,698	0,717	0,45	0,505	0,527	0,544	0,572
	0,3	NONE	0,592	0,643	0,662	0,675	0,3	0,455	0,486	0,511	0,531	0,559
	0,3	RUSBoost	0,603	0,655	0,675	0,697	0,703	0,431	0,553	0,558	0,606	0,626
	0,3	SMOTE	0,636	0,679	0,687	0,704	0,713	0,489	0,549	0,562	0,585	0,598
	0,3	SMOTEBoost	0,641	0,678	0,687	0,704	0,713	0,476	0,55	0,561	0,585	0,598
RF	0,1	NONE	0,531	0,573	0,608	0,099	0,67	0,36	0,309	0,335	0,405	0,473
	0,1	RUSBoost	0,715	0,765	0,806	0,827	0,842	0,537	0,64	0,67	0,683	0,707
	0,1	SMOTE	0,631	0,684	0,692	0,728	0,741	0,448	0,45	0,467	0,525	0,552
	0,1	SMOTEBoost	0,627	0,669	0,684	0,721	0,734	0,376	0,412	0,453	0,516	0,542
	0,15	NONE	0,138	0,146	0,148	0,149	0,685	0,34	0,334	0,395	0,46	0,509
	0,15	RUSBoost	0,693	0,76	0,768	0,8	0,807	0,499	0,659	0,654	0,687	0,694
	0,15	SMOTE	0,64	0,686	0,682	0,734	0,747	0,412	0,476	0,474	0,558	0,574
	0,15	SMOTEBoost	0,644	0,673	0,678	0,731	0,741	0,393	0,448	0,467	0,551	0,566
	0,25	NONE	0,607	0,247	0,674	0,249	0,706	0,374	0,449	0,508	0,551	0,57
	0,25	RUSBoost	0,692	0,71	0,747	0,761	0,773	0,613	0,621	0,654	0,681	0,685
	0,25	SMOTE	0,684	0,7	0,732	0,736	0,753	0,514	0,545	0,593	0,602	0,625
	0,25	SMOTEBoost	0,679	0,703	0,733	0,736	0,752	0,509	0,542	0,59	0,598	0,62
	0,3	NONE	0,623	0,666	0,69	0,707	0,3	0,424	0,505	0,554	0,587	0,607
	0,3	RUSBoost	0,672	0,715	0,738	0,751	0,764	0,579	0,638	0,649	0,674	0,682
	0,3	SMOTE	0,685	0,715	0,734	0,744	0,756	0,546	0,597	0,624	0,638	0,653
	0,3	SMOTEBoost	0,684	0,717	0,737	0,745	0,757	0,539	0,596	0,623	0,635	0,651

CM: Classification Methods

5.2. Real Datasets Results

The results for real data sets were given in Table 7. According to the results, Balanced Accuracy was higher than F-Measure in each dataset. When looking at the formula for the Balanced Accuracy measure, it is only calculated from the values of sensitivity and specificity, not affected by prevalence. The F-measure value takes into account the prevalence. Therefore, we considered the F-measure when interpreting the results.

The effect on the classification accuracy of the algorithms used is clearly seen in Table 7. The effect on the classification results of the algorithms used in datasets with high imbalanced was found to be similar. As the imbalance rate decreases, RUSBoost's effect increases comparing to other techniques (SMOTE and SMOTEBoost).

Table 7: Real Datasets Results.

Datasets	Algorithms	SVM		CART		RF	
		Balanced Accuracy	**F-Measure**	**Balanced Accuracy**	**F-Measure**	**Balanced Accuracy**	**F-Measure**
Abalone	NONE	0.500	NA	0.581	0.267	0.623	0.375
	SMOTE	0.583	0.286	0.578	0.250	0.576	0.375
	SMOTEBoost	0.583	0.286	0.655	0.250	0.662	0.235
	RUSBoost	0.812	0.286	0.709	0.250	0.682	0.235
Fertility	NONE	0.500	NA	0.500	NA	0.667	0.500
	SMOTE	0.423	NA	0.404	NA	0.404	NA
	SMOTEBoost	0.647	0.400	0.500	0.182	0.647	0.500
	RUSBoost	0.500	0.400	0.500	0.182	0.622	0.500
Thoracic Surgery	NONE	0.500	NA	0.500	NA	0.500	NA
	SMOTE	0.451	0.113	0.547	0.268	0.514	0.248
	SMOTEBoost	0.416	0.113	0.520	0.268	0.533	0.269
	RUSBoost	0.539	0.113	0.551	0.261	0.548	0.269
Hepatitis	NONE	0.500	NA	0.500	NA	0.767	0.667
	SMOTE	0.833	0.800	0.642	0.400	0.895	0.667
	SMOTEBoost	0.833	0.800	0.642	0.400	0.921	0.667
	RUSBoost	0.733	0.800	0.658	0.400	0.796	0.600
Blood Transfusion	NONE	0.625	0.410	0.500	NA	0.615	0.404
	SMOTE	0.553	0.247	0.540	0.219	0.593	0.366
	SMOTEBoost	0.553	0.247	0.540	0.219	0.597	0.366
	RUSBoost	0.730	0.477	0.666	0.505	0.621	0.582
Alzheimer	NONE	0.560	0.364	0.476	0.200	0.595	0.400
	SMOTE	0.679	0.571	0.679	0.571	0.679	0.400
	SMOTEBoost	0.679	0.571	0.679	0.571	0.679	0.571
	RUSBoost	0.631	0.571	0.536	0.571	0.488	0.571
Pima Diabetes	NONE	0.667	0.517	0.673	0.551	0.679	0.548
	SMOTE	0.731	0.641	0.712	0.613	0.763	0.683
	SMOTEBoost	0.731	0.641	0.673	0.613	0.750	0.683
	RUSBoost	0.744	0.641	0.712	0.613	0.788	0.683

NA (Not Available): The F-measure value could not be calculated because the sensitivity values were zero.

6. Discussion

In this research, the data sets obtained by the simulation study and the real data sets were studied. The results of different classification methods in two-class datasets were evaluated on the Balanced Accuracy and F-measure used as performance measures. In most studies in the literature, algorithms have been evaluated on real data sets. Therefore, there is no study similar to the simulations discussed in this study in the literature. In 2002, Chawla and his colleagues [1] they tried to increase the number of minority class observations by producing new (artificial) observations on the minority class by following a different approach with the SMOTE algorithm. On the other hand, in 2010 Seiffert and his colleagues [6] showed that drawing samples from the majority class to the minority class with the RUSBoost algorithm gives better results. In our simulation study, the RUSBoost algorithm is among the algorithms that give good results in 4 different correlation structures. In addition, Seiffert and his colleagues [6] showed that RUSBoost gives results in a short time in the background according to SMOTE and SMOTEBoost algorithms.

7. Conclusion

The problem of class imbalance is an important issue in machine learning. There are many studies on this subject. The number of researches is increasing day by day and new algorithms are proposed. A few of them have worked both of real datasets and simulation, however, these simulation studies were not examined under correlation structures. A simulation study was carried out with different sample size, correlation structure and class imbalance ratios from these studies. These effects affect the accuracy of classification. We tried to show the validity of algorithms which are using for class imbalance problem in a simulation study.

As a result, if algorithms were not used, they would not be able to correctly classify them, so they would be misclassified. Also, accuracy of classification was significantly lower. According to simulation and real dataset's results, imbalanced was decreased or eliminated and performance measures increased. Accordingly, it is necessary to look at the imbalance at the beginning of the study and use data balancing algorithms if there is an imbalance.

There was no highly correlated data in real dataset, high correlation data in real datasets could also be included in the study. Studies can be extended for multiclass datasets and algorithms can be applied to gene datasets as well. In our simulation study, the relationships between independent variables were evaluated as low, medium and high. Apart from this, the relationship structures that are not taken into account can also be examined.

References

[1] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, ”SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artifical Intelligence Research*, vol. 16, pp. 321-357, 2002.

[2] H. He, Y. Ma, *Imbalanced Learning.* Hoboken, New Jersey, pp. 13-36, 2013.

[3] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches,” *IEEE Trans. On Systems, Man, And Cybernetics—Part C: Application And Reviews*, vol. 42, no. 4, July 2012.

[4] X. Liu, J. Wu, and Z. Zhou, “Exploratory Undersampling for Class-Imbalance Learning,” *IEEE Trans. On Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539-550, December 2008.

[5] N.V. Chawla, A. Lazarevic, L.O. Hall and K.W. Bowyer, “SMOTEBoost: Improving Prediction of the Minority Class in Boosting,” *Proc. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 107-119, 2003, doi:10.1007/978-3-540-39804-2_12.

[6] C. Seiffert, T.M. Khoshgoftaar, J.V. Hulse and A. Napolitano,”RUSBoost: A Hybrid Approach to Alleviating Class Imbalance,” *IEEE Trans. On Systems, Man, And Cybernetics—Part A: System and Humans,* vol. 40, no. 1, pp. 185-197, Jan 2010, doi: 10.1109/TSMCA.2009.2029559.

[7] A. Estabrooks, T. Jo And N. Japkowicz,” A Multiple Resampling Method for Learning from Imbalanced Data Sets,” *Computational Intelligence,* vol. 20, no.1, pp. 18-36, Feb 2004, doi: 10.1111/j.0824-7935.2004.t01-1-00228.x

[8] C.V. KrishnaVeni, T. Sobha Rani,”On the Classification of Imbalanced Datasets,” *International Journal of Computer Science and Technology* vol. 2, no. 1, pp. 145-148, December 2011.

[9] J. Zhang and I. Mani, “kNN approach to unbalanced data distributions: A case study involving information extraction,” *Proc. Int. Conf. Mach. Learning*, pp. 42–48, 2003.

[10] P. Cao, D. Zhao, and O. Zaiane,”An optimized cost-sensitive SVM for imbalanced data learning,” *Proc. Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining,* pp. 280-292, 2013.

[11] D. Cieslak and N. Chawla, “Learning decision trees for unbalanced data,” *Proc. Machine Learning and Knowledge Discovery in Databases*, pp. 241–256, 2008.

[12] N. V. Chawla, N. Japkowicz, and A. Kotcz, Eds.,”Imbalanced Data Sets*,” Proc. ICML Workshop Learn*, 2003.

[13] C. Drummond and R. C. Holte,”C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling,” Proc. *Workshop on Learning from Imbalanced Data Sets II, International Conference on Machine Learning*, 2003.

[14] N. Zhong, L. Zhou, Springer Verlag, *Methodologies for Knowledge Discovery and Data Mining*, Third Pacific-Asia Conference, Pakdd-99, (1999).””,

[15] V.N. Vapnik, *The Nature of Statistical Learning Theory*. Verlag: New York., pp. 123-167, 1999.

[16] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco: CA, pp. 285-382, 2006.

[17] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. pp. 143-151, 2005.

[18] H. Demirtas , A. Amatya and B. Doganay,” BinNor: An R Package for Concurrent Generation of Binary and Normal Data,” *Communications in Statistics - Simulation and Computation*, vol. 43, no. 3, pp. 569-579, Jan 2014, doi: 10.1080/03610918.2012.707725.

[19] M. Kuhn,”Building Predictive Models in R Using the caret Package,” *Journal of Statistical Software*, vol. 28, no. 5, pp. 1-26, 2008, doi: 10.18637/jss.v028.i05

[20] T. Therneau, B. Atkinson and B. Ripley, https://cran.r-project.org/web/packages/rpart/rpart.pdf. 2014.

N. Lunardon, G. Menardi, and N. Torelli,” ROSE: A Package for Binary Imbalanced Learning,” *The R Journal*, vol. 6, no. 1, 2014.