Araştırma Makalesi / Research Article

Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data

Guhdar A. A. MULLA[1], Yıldırım DEMİR[2*], Masoud M. HASSAN[3]

[1]*Van Yuzuncu Yil University, Institute of Natural and Applied Sciences, Department of Statistics, Van, Turkey*
[2]*Van Yuzuncu Yil University, Faculty of Economics and Administrative Sciences, Department of Statistics, Van, Turkey*
[3]*Zakho University, Faculty of Science, Department of Computer Science, Zakho, Kurdistan Region – F.R. Iraq*
(ORCID: 0000-0001-6742-0083) (ORCID: 0000-0002-6350-8122) (ORCID: 0000-0003-3461-0942)

Abstract
Imbalanced data classification is a common issue in data mining where the classifiers are skewed towards the larger data class. Classification of high-dimensional skewed (imbalanced) data is of great interest to decision-makers as it is more difficult to. The dimension reduction method, a process in which variables are reduced, allows high dimensional datasets to be interpreted more easily with a certain loss. This study, a method combining SMOTE oversampling with principal component analysis is proposed to solve the imbalance problem in high dimensional data. Three classification algorithms consisting of Logistic Regression, K-Nearest Neighbor, Decision Tree methods and two separate datasets were utilized to evaluate the suggested method's efficacy and determine the classifiers' performance. Respectively, raw datasets, converted datasets by PCA, SMOTE and SMOTE+PCA (SMOTE and PCA) methods, were analyzed with the given algorithms. Analyzes were made using WEKA. Analysis results suggest that almost all classification algorithms improve their classification performance using PCA, SOMTE, and SMOTE+PCA methods. However, the SMOTE method gave more efficient results than PCA and PCA+SMOTE methods for data rebalancing. Experimental results also suggest that K-Nearest Neighbor classifier provided higher classification performance compared to other algorithms.

Keywords: Classification, Dimensionality reduction, Imbalanced classes, PCA, SMOTE oversampling.

Yüksek Boyutlu Dengesiz Verilerin Sınıflandırılması İçin SMOTE Aşırı Örnekleme İle PCA'nın Kombinasyonu

Öz
Dengesiz veri sınıflandırması, sınıflandırıcıların daha büyük veri sınıfına doğru çarpıtıldığı veri madenciliğinde yaygın bir konudur. Yüksek boyutlu çarpık (dengesiz) verilerin sınıflandırılması, daha zor olduğundan karar vericiler için büyük ilgi görmektedir. Değişkenlerin azaltıldığı bir süreç olan boyut küçültme yöntemi, yüksek boyutlu veri setlerinin belirli bir kayıpla daha kolay yorumlanmasına olanak tanır. Bu çalışmada, yüksek boyutlu verilerdeki dengesizlik problemini çözmek için SMOTE aşırı örneklemeyi temel bileşen analizi ile birleştiren bir yöntem önerilmiştir. Önerilen yöntemin etkinliğini değerlendirmek ve sınıflandırıcıların performansını belirlemek için Lojistik Regresyon, K-En Yakın Komşu, Karar Ağacı yöntemlerinden oluşan üç sınıflandırma algoritması ve iki ayrı veri kümesi kullanılmıştır. Sırasıyla, ham veri setleri, PCA, SMOTE ve SMOTE +PCA (SMOTE ve PCA) yöntemleriyle dönüştürülen veri setleri, verilen algoritmalarla analiz edilmiştir. Analizler WEKA ile yapılmıştır. Analiz sonuçları, neredeyse tüm sınıflandırma algoritmalarının PCA, SOMTE ve SMOTE+PCA yöntemlerini kullanarak sınıflandırma performanslarını iyileştirdiğini göstermektedir. Bununla birlikte, SMOTE yöntemi, verilerin yeniden dengelenmesi için PCA ve PCA+SMOTE yöntemlerinden daha verimli sonuçlar vermiştir. Deneysel sonuçlar ayrıca K-En Yakın Komşu sınıflandırıcısının diğer algoritmalara kıyasla daha yüksek sınıflandırma performansı sağladığını göstermektedir.

Anahtar kelimeler: Sınıflandırma, Boyut azaltma, Dengesiz sınıflar, PCA, SMOTE aşırı örnekleme.

[*]Corresponding author: ydemir@yyu.edu.tr
Received: 20.05.2021, Accepted: 28.07.2021

1. Introduction

One of the sciences developed to examine the phenomena in nature and solve existing or potential problems is data analysis science. Data analysis science aims to explain the subject with a certain probability by using methods and theories suitable for data structure with a limited number of observations and shed light on future research. There are many data in every science field in nature, and access to these data is getting easier day by day. However, how much of the obtained data can be used or how important it always comes to the fore. In addition to obtaining information, mankind beings consume data even in their daily work and produce more than they consume.

Due to the intensive use of data, it is necessary to classify the data to make it useful and generate new information. Classification is the grouping of a product or data according to determined distinguishing features through algorithms. It is often impossible to manually classify data. Because millions of data types are formed related to a field even in just one day, at the same time, subjecting data to many algorithms to classify, it can produce different results. The effective classification depends on which algorithm will be applied to the dataset. Thus, it is likely that not every algorithm will give the same accuracy to all datasets and that the algorithms used to model the data are too low over the dataset. In this direction, machine learning and classification algorithms come to the fore [1].

There are several different algorithms for classification that exist in the literature. Decision Tree (DT), K-Nearest Neighbor (K-NN), and Logistic Regression (LR) methods are the most important and the most common machine learning algorithms for the classification process. With experimental training data and the latest test data, all of these sophisticated classification algorithms usually have a high accuracy rate. Given the high accuracy pattern, model developers could find it challenging to develop a better classification system. In addition, the provision of such high predictive precision may mean that machine learning techniques can correctly solve almost any classification problem. However, in solving every problem, such high prediction accuracy cannot be seen [2].

One of the most difficult issues in the classification algorithms is the classification of imbalanced data. Because a problem occurs in binary classification when the sample sizes are not equal in both classes, in other words, one class has many samples called majority, while the other class has relatively few samples called minority. However, this problem may not be very important if there is very little difference between samples from the positive and negative groups. In addition, when data are imbalanced, the majority class usually considers the key features of interest to learn from while ignoring the impact of the minority in the dataset. Nonetheless, most conventional issues cannot be solved by classification algorithms, as while they are designed to achieve high overall precision, they are most likely to misclassify positive class samples, which is a disadvantage of classification under imbalanced data. To find a good answer with good precision for both the positive and negative groups has become a significant research area [3]. In order to classify imbalanced data, different oversampling or under-sampling should be used first. SMOTE (Synthetic Minority Oversampling Techniques) is a technique for rebalancing a dataset. This approach provides an optimal solution to the unequal data delivery problem caused by oversampling. [4].

Another classification issue is the high dimensionality of data, where there are lots of redundant features in the dataset. Using dimension reduction methods is to reduce the irrelevant features from the data before applying any classification algorithms. A variety of data processing goals depend on the reduction of measurements. Input selection in classification problems is a mission-specific dimension reduction form. High-dimensional data visualization involves mapping to a lower dimension, usually three or less. Principal Component Analysis (PCA) is a very-well known classic method for linear dimension reduction. One performs an orthogonal transformation to correlate vectors and projects spanning the subspace corresponding to the highest eigenvalues by such eigenvectors in PCA. This conversion makes the signal unrelated components, and the projection along the high-variance directions maximizes variance while minimizing the mean squared residual between the initial signal and its dimension reduction approach [5].

This study aimed to find a solution to the sorting issue for high-dimensional, imbalanced data. In the study, the classification problem is investigated (using the PCA and SMOTE methods) by reducing unnecessary features with rebalancing simultaneously. Following this purpose, for separate imbalanced datasets where the number of samples in one group (a large percentage) is significantly higher than the number of specimens in the other class, three well-known classification algorithms (DT,

K-NN, and LR) were used (minority). According to the other majority classes, the amount of data in the minority classes is insufficient to obtain adequate information in the extremely imbalanced datasets. Weka program was used in data analysis. Thanks to this program, the data have classified by rebalancing and reducing its dimension [6].

2. Material and Method

Two separate datasets were used to evaluate the feasibility and efficiency of the proposed approaches in this research. These datasets consisted of heart disease attack and lymphography-normal-fibrosis data and were obtained from addresses https://www.kaggle.com/johnsmith88/heart-disease-dataset and https://sci2s .ugr.es/keel/imbalanced.php#sub20 on 12.06.2020, respectively. Furthermore, the first dataset consists of 14 variables and 1025 data (499 negative class and 526 positive class), while the second dataset consists of 19 variables, 148 data (142 negative and 6 positive). Thus, the first dataset's imbalance rate is 2.6%, while the second is 92%. The analyzes were made with the WEKA program.

2.1. Related Literature

Classification is a significant pattern recognition activity. Many classification learning algorithms have been well evolved and successfully applied to many application domains, including decision trees, k-nearest neighbour, logistic regression, and the recently published associative grouping. The unequal class distribution of a dataset, on the other hand, has proven to be a significant challenge for most classifier learning algorithms, implying a reasonable [7].

Lorena et al. (2019), examined resampling methods and metrics that could be derived from training datasets in order to describe the difficulty of the classification problems in this paper. Their methods were also examined and discussed in recent literature, allowing for potential research opportunities in the field. Finally, definitions were given for the Extended Complexity Library (ECoL) R package, which outfits a series of complication measures and has been completed publicly accessible.

Basgall et al. (2018) paper introduce SMOTE-BD, for imbalanced sorting in Big Data, a completely scalable preprocessing method has been created. It was centered on the SMOTE algorithm, which generates new synthetic samples based on the proximity of each minority class instance, and is one of the most commonly employed preprocessing solutions for unequal categorization Their novel production was made to be self-contained.

Mohammed et al. (2020), for a new real diabetes dataset, researchers proposed a method to study, diagnose, and identify imbalanced diabetes patients using six machine learning algorithms. The new dataset, dubbed ZADA, was compiled from the medical records of approximately 7000 patients in the Iraqi city of Zakho. To address the issue of class imbalance, they suggested a classification analysis based on the three normalization methods and the resampling (SMOTE) process. Various studies were carried out to find the best algorithm with the best results based on minority class distribution. According to the findings, the resampling process and normalization techniques positively impacted the classification model performance.

2.2. Classification Algorithms

Classification is the method of deciding which semantic groups those objects belong to depending on their features [10]. It is of assigning new observed samples by examining the sample's features to an existing defined class. It can then decide the unseen cases by constructing a model from previous data [11]. It uses data mining techniques to investigate the relationship between each row's variables' values and the label given to that row. There are several different classification algorithms, such as LR, K-NN and DT. These algorithms use different representations of these relationships so that when new rows are fed to the classifier, the extracted knowledge can be applied to predict a label for that row, depending on the variable values that characterize the row. These relationships differ from one dataset to another, so it is critical that the classifiers are trained using a labeled training dataset [12]. For example, a Medical Database The training set must have previously reported relevant patient information; Whether or not the patient has a heart attack is the prediction trait. The overall theme beyond Data Mining categorization is to determine the training dataset's target class [13].

2.2.1. Logistic regression

Logistic Regression (LR) operates very much like linear regression but with a variable binomial response. This algorithm's main advantage is that you can use continuous explanatory variables, and it is simpler to use them. Simultaneously treat more than two explanatory variables. The LRA is a vector with one or more explanatory variables that can be used to investigate the implicit relationship between two reactions. The logistic variable model forecasts the X Logit Y representing a normal chances logarithm of Y for one explanatory event, the X variable with one Y binary outcome variable. The main formula of the logistic regression model can be defined as follows [14], [15].

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \tag{1}$$

The left side is referred to as the log-odds or logit. The logit in the LR model is linear in X.

$$\text{Consequently: } \pi(X) = E\left(\frac{Y}{X}\right) = \frac{e^{a+BX}}{1 + e^{a+BX}} \tag{2}$$

π Where π is the likelihood of the desired outcome, expressed as $X = x_1, x_2, \dots, x_k$, α is a parameter representing the *Y*-intercept, and β is a parameter of the slope, X can be qualitative (categorical) or quantitative variables, and Y is either measurable or categorical at all times. From basic to multiple linear regression, the equation (3) could be represented and generalized as follows:

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{3}$$

Therefore,

$$\pi(x) = \frac{e^{a+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1 + e^{a+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \tag{4}$$

2.2.2. K-Nearest neighbor

One of the most important and straightforward grouping methods is K-Nearest Neighbor (K-NN). When there is virtually no prior information on the knowledge appropriation, it can be one of the most important decisions to analyze a classification [16].

The Euclidean distance between the defined training specimens and a reference sample is the basis for the K-NN classifier. Let x_i be an input sample with m features $(x_{i1}, x_{i2}, \dots, x_{in})$, n is the cumulative number of specimens in the input $(i = 1, 2, \dots, n)$ and m the total number of features [17]. The Euclidean distance between sample x_i and $x_j (j = 1, 2, \dots, n)$ is defined;

$$\mathrm{D}(x_i, x_j) = \sqrt{(x_1 - x_1)^2 + (x_2\ x_2)^2 + \cdots + (x_n - x_n)^2} \tag{5}$$

Where x_i and x_j are subjects to be compared with n characteristics. There are also other methods to calculate distance, such as Manhattan distance.

The value of k in the K-NN algorithms can be preferred for neighbors. The appropriate choice of k has an important influence on the K-NN algorithm for diagnostic results. A big value of k reduces the effect of random error-induced variance; however, there is a possibility that small but meaningful trends can be overlooked. The secret to choosing the correct k value is to achieve a balance from overfitting to under fitting [18].

2.2.3. Decision trees

Decision Trees (DT) are classification trees that sort samples according to the function's values. Starting from the root node, each node in a decision tree reflects a function in the specimens to be categorized, and each branch determines a value that the node can infer; samples are clustered and classified depending on their values. Decision trees are a statistical algorithm that uses a decision tree to map assumptions about an entity to the item's target value. Since decision trees are pruned utilizing a validation system, most decision tree classifiers use post-pruning techniques to test their performance. Per node may be removed and reassigned to the most popular training sample category [19].

A decision tree reflects the learned function of decision tree learning, which resembles discrete-valued target functions. These learning approaches are some of the most widely used inductive inference algorithms, and they have been effectively extended to a wide series of learning activities. Hospital training cases are diagnosed to assess the creditworthiness of loan applicants. [20].

Using a Decision Tree, we start by defining entropy, a normally used measure in information theory, to specifically identify information and gain. Provided the *S* list, which includes positive the entropy of (S) relative to this, and negative examples of a certain target term, Boolean scoring is defined as follows

$$Entropy\,(S) = -\sum_{i=1}^{n} P_i\, log_2(P_i) \qquad i: 1, 2, \dots, n \tag{6}$$

where S: Let *S* be a resource, P: A probability distribution and n: Simple Volume

So far, we've only looked at entropy in the context of Boolean target classification. If the target variable can have c different values, the entropy of S concerning this c-wise classification is known as

$$Entropy\,(s) = \sum_{i=1}^{c} -p_i\, log_2\, p_i \tag{7}$$

where p_i denotes the percentage of S that belongs to class i. Since entropy is a measure of predicted encoding length measured in bits, the logarithm is still in base 2. It's also worth noting that if the target variable has C possible values, the entropy will reach $logC$.

2.3. Principal Component Analysis (PCA)

PCA is a dimensionality reduction process that employs an orthogonal transformation to convert a collection of potentially linked variables into a set of linearly uncorrelated variables known as the principal component. The original variable number is less than or equal to the principal component number. Every corresponding element has the highest possible variation under the condition that the previous component is orthogonal, and the first key factor has the most potential for variation (i.e. accounts for as much data uncertainty as possible). The main components are orthogonal, as they are the own vectors of the symmetric covariance matrix. The relative scaling of the original variables is subject to the PCA [21].

When applying the PCA, the random variables' variance is shown by the covariance matrix's eigenvalue that is located at the main diagonal of a matrix $\sum$. In matrix $\sum$ the eigenvalues are sorted according to their magnitudes, from bigger to smaller. That means the PCA with bigger variance comes first. Table 1 shows the general shape of high dimensional data with N samples and m variables.

Table 1. The shape of a dataset consisting of *n* samples, each has m variable

		Variables			
		a_1	a_2	...	a_m
Samples	x_1	a_{11}	a_{12}	...	a_{1m}
	x_2	a_{21}	a_{22}	...	a_{2m}
	⋮	⋮	⋮	⋱	⋮
	x_N	a_{N1}	a_{N2}	...	a_{Nm}

The main purpose of using the PCA is to reduce the high-dimensional data with a dimension m into a lower-dimensional data of dimension k, where$k \leq m$.

1. Let $x_1, x_2, \ldots, x_m$ be *m* continuous predictors. The following is a summary of fundamental element analysis input C_{mxm}, the covariance matrix of $x_1, x_2, \ldots, x_m$.
2. Calculate the eigenvectors and eigenvectors of the covariance matrix. Sort the eigenvalues (and corresponding) in descending order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$.
3. Come up with new predictors. Assume the first component's aspects v_1 are $v_{11}, v_{22}, \ldots, v_{1m}$, then the new derived predictors are $\frac{v_{11}}{\sqrt{\lambda_1}} x_1 + \frac{v_{12}}{\sqrt{\lambda_1}} x_2 + \cdots + \frac{v_{1m}}{\sqrt{\lambda_1}} x_m$.

2.4. Class Imbalance Problem

Through the data mining process' pattern extraction stage, learning algorithms are commonly used. Since this is real-world data, there have been several difficulties in implementing current and well-learning algorithms. Numerous learning algorithms are based on the premise of evenly distributed class distributions or no significant variations in class prior probabilities. Nevertheless, in real-world data, one class may not always be represented by many instances, while a few may only describe the other.

According to several studies, there are 1% class inequalities in the minority class and 99 percent and higher in the majority class. Learning algorithms in these situations attempt to generate classifiers with incredibly low overall error rates by categorizing each conceptual model as contributing to the majority class. These classifiers are ineffective because the minority class with rare cases is most concerned with an accurate prediction. Several concepts extend an algorithmic approach to class inequality by adapting current algorithms and strategies to skewed results' unique features. Such principles include cost-sensitive preparation, one-class classificatory, and classifier ensembles. The goal of cost-sensitive learning is to lower misclassification costs. Class inequality may be handled similarly by allocating higher classification expenses to classes identified by just a few examples [22].

There are different methods to solve the imbalance problem, and the most common one is the SMOTE method.

2.4.1. SMOTE method

The Synthetic Minority Over-Sampling Technique (SMOTE) is a synthetic data generation over-sampling technique. Its key concept is to interpolate between many examples of minority classes lying together to generate new examples of minority classes. SMOTE, to be more specific, chooses a group, E_i, and its neighbors at random. A new example is generated using the equation below, using an example E_i from the nearest neighbor set:

$$Enew = E_i + (E_i - E_i)\delta \quad (8)$$

Where δ is fixed in the interval selected at random (0,1). SMOTE could enlarge the minority class's space by enabling the development of synthetic instances that stretch deeper into the prevailing class's space so the preferred closest neighbor could be like a class other than E_i [23].

A basic example of SMOTE method is shown in Figure 1. The sample x_i from the minority, class is chosen as the basis for generating new synthetic data points. Focused on the distance metric is selected, with several nearest neighbors of the same class (points x_{i1} to x_{i4}). Finally, simultaneous interpolation is used to generate new specimens r_1 to r_4.

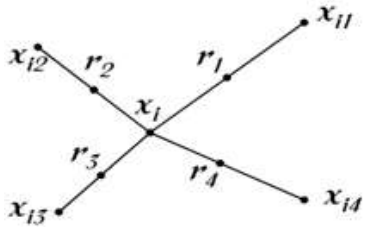

Figure 1. An example of how the SMOTE algorithm generates synthetic data points

2.5. Design of the Proposed Classification Model

Classification of high-dimensional data with class imbalance problem is of great interest to researchers in different science fields. Dealing with these two problems (high-dimensionality and class imbalance) for classification algorithms is challenging yet very beneficial. This thesis proposed a new method based on combining two well-known techniques: (PCA for dimensionality reduction) and (SMOTE for rebalancing) to tackle these two problems together. Figure 2 shows the proposed method's diagram to choose the best classification model, compare classification results before and after using dimensionality reduction by PCA as dimensionality reduction, compare the classification results, and rebalance the dataset the SMOTE oversampling method. The method starts with using the original dataset, and then we will check if it is imbalanced. If the dataset is imbalanced, then we will balance it by using the SMOTE method. After this process, we will use PCA to reduce the dimensionality of the data. But if the dataset is already balanced then we will directly use the PCA. After the data have been rebalanced and their dimensionality is reduced, we will split the dataset into training and testing to find out the training percentage, which is 66%, along with the testing percentage, which is 34%. By the end of this process, we will classify the dataset using the three classification algorithms which are (LR, K-NN and DT). Thus, we will evaluate all three classification models using their evaluating measurements: Accuracy, F-measure, ROC area (the area under the ROC curve). We used ten-fold cross-validation for evaluation: the knowledge was automatically split into ten pieces of equal size, and the training and testing procedure was carried out ten times, each component being the test data once and the remaining pieces for training. After preparing the data for classification and the assessment, our method gives each test record to the most likely class. The next and last step is to compare them separately by comparing classification performance before and after using the PCA, SMOT and combined SMOT with PCA according to the evaluating of the classification models to determine if the classification models' accuracy has been improved or not.

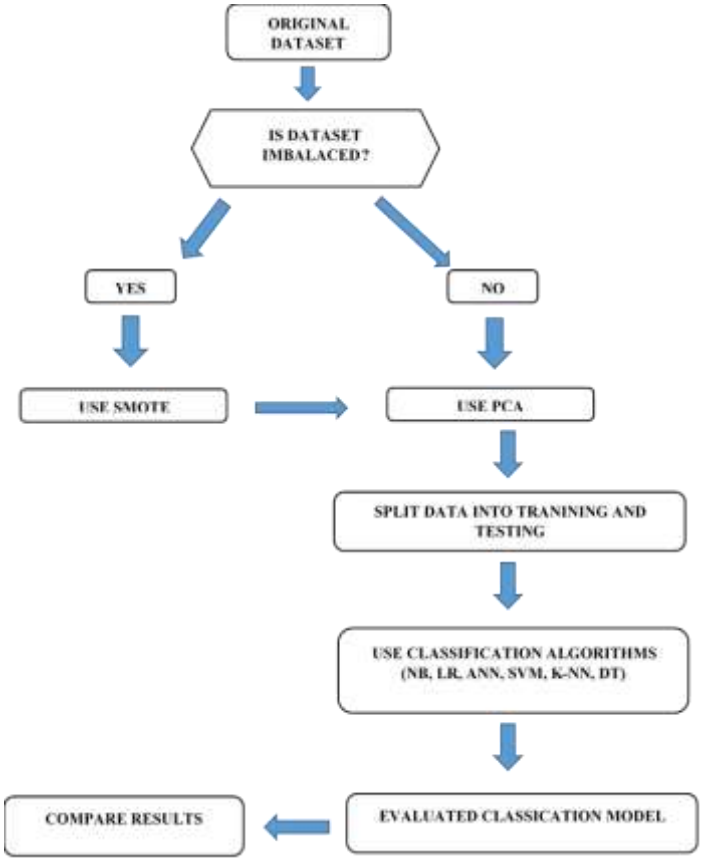

Figure 2. Design of the proposed classification model

To check each classification model's efficiency and performance, we use the confusion matrix (defined below) to calculate different evaluation metrics as follows.

Accuracy

Classification data mining algorithms have been suggested as a fitness method to evaluate and subset generated in this analysis. The following is the method for calculating the precision of each subset:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

Where false negative (FN) denotes a positive sample that has been wrongly categorized as negative, true positive (TP) means that a promising case was properly classified, false positive (FP) denotes the incorrect classification of a negative example as positive, and true negative (TN) denotes the correct classification of a negative example as negative.

F-Measure

You can't skip the other metric, F- Measure, that is Precision and Recall's function, if you read a lot of other research on Precision and Recall. The following is the formula:

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

When you need to find a balance between accuracy and memory, this is the tool to use, you'll need F-measure. So, what exactly is the distinction between F-measure and Accuracy? We've shown that a large number of True Negatives will contribute significantly to accuracy, which we don't focus on too often in most business situations, while False Negative and False Positive typically have business costs (tangible and intangible), since we need to consider a balance between Precision and Recall and there is an unequal class distribution, the F-measure could be a safer test to use (large number of Actual Negatives).

ROC Area

An ROC region (receiver operating characteristics curve) is a graph that displays the efficiency of a classification system over all classification levels True Positive Rate (*TPR*) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (*FPR*) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots *TPR* vs. *FPR* at diverse classification thresholds. As the rating criterion is lowered, more objects are classified as positive, growing both False Positives and True Positives.

3. Results and Discussion

During this study, there were 4 different methods have been applied to the data under study. The four methods include: (1) classification of raw data, (2) classification with the use of PCA, (3) classification with the use of the SMOTE, and (4) classification with the use of the combination of both, SMOTE and PCA. Each of these methods was applied using all three algorithms: LR, K-NN, and DT. When we used LR algorithms, we set the ridge value in the log-likelihood was 1.0E-8 and the dataset was divided into 10 folds for cross-validations. For the K-NN algorithms we used k=1, and the dataset divided to 10 cross-validations. When we used DT algorithms, the confidence factor used for pruning was 0.25 and determined the amount of data used for reduced-error purring 3. The first dataset was reduced to 7 ranker variables when using the PCA method, while the second dataset was reduced to 10 ranker variables. The

variance covered that we used in PCA method was 0.95. On the other hand, when we applied the SMOTE method, we used k=5 and the seed used for random sampling was equal to 1, the percentage of SMOTE sample to create was equal to 100%.

The precision, F-measure, and ROC region measurements were measured for testing the efficiency of the three classification algorithms prior even after using dimensionality reduction and rebalancing data techniques, as shown in Table 2.

Table 2. Comparing the performance of the classification algorithms using for dataset 1

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
LR	79.7	79.7	89.7	79.7	79.7	89.7	99.4	99.4	99.2	83.2	83.1	91.2
K-NN	**100**	**100**	**100**	**99.7**	**99.7**	**99.3**	**99.8**	**99.8**	**100**	**99.5**	**99.5**	**99.6**
DT	83.9	83.8	88.9	85.1	85	90.1	96.1	96.2	97.8	97.9	97.9	98.9

According to Table 2, for the raw data, the highest Accuracy, F-measure and ROC area rate of 100%, 100% and 100% were obtained in K-NN, respectively; According to the three measurements calculated, the lowest rates as 79.7%, 79.7% and 89.7% were obtained in LR respectively. However, when the PCA dimensionality reduction method was used, we can see that the highest Accuracy and F-measure ROC area rates of 99.7%,99.7%, and 99.3% were obtained K-NN, respectively. When the SMOTE oversampling method was applied, the highest Accuracy, F-measure and ROC rate of 99.8% 99.8%, 100% were obtained in K-NN, respectively. On the other hand, when the two methods (PCA + SMOTE) were simultaneously applied, we can see that the classification performances were further improved for almost all the classifications algorithms used, and the highest accuracy, F-measure and ROC rates of 99.5%, 99.5%, and 99.6% were obtained in K-NN algorithm, respectively. Looking at Table 2 results in more details, we can see that after reducing the dimensionality and rebalancing data, the performances of the classification algorithms have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data under the imbalance problem.

Figure 3 shows each classification algorithms' performance for Dataset 1, before and after using PCA, SMOTE, and PCA+SMOTE using Accuracy (Acc.), F-measure (F) and ROC metrics.

Figure 3. Performance of each classification algorithms for Dataset 1

According to the shape of the LR algorithm given in Figure 3, it is obvious that there is a higher improvement of Accuracy, ROC area and F-measurement methods when using SMOTE method. The results of the K-NN algorithm figure showed that after using all three methods, the results were worst, and this is because this algorithm has already provided optimal results that do not require any further improvement. When applying the DT algorithms with the three methods we used, the classification performance of almost all the classifiers was better with either one of the three methods.

Table 3 shows the performance of the three classification algorithms for Dataset 2, before and after using PCA, SMOTE, and PCA+SMOTE methods to deal with the imbalance problem with high-dimensional data.

Table 3. Comparing the performance of the classification algorithms using for dataset 2

Algorithms	Raw Data			PCA			SMOTE			PCA + SMOTE		
	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC	Acc.	F	ROC
LR	97.2	97.5	**97.8**	98.6	98.6	**99.4**	98	98.1	96.6	98.7	98.8	**99.9**
K-NN	**98.6**	**98.5**	89	**99.3**	**99.3**	93.3	**98.7**	**98.6**	**97.2**	**99.3**	**99.3**	98.3
DT	97.9	97.6	54.7	97.2	97.5	91.2	96.1	95.9	83	98.7	98.8	99.2

According to Table 3, for the raw data, the highest Accuracy and F-measure rate of 98.5% and 98.5 were obtained in K-NN respectively, but the highest ROC area performance rate of 97.8% was obtained in LR. When using the PCA method, K-NN and LR have shown significant improvement in classification performance, while this method has not provided effective DT effectiveness. With the SMOTE method, LR and K-NN classifiers provided better results than DT classifiers and the DT gave the worst results. However, when the combination of (PCA+SMOTE) is used, all the classification algorithms have improved the evaluation metrics used. Looking at Table 3 results in more details, we can see that after reducing the dimensionality and rebalancing data, the classification algorithms' performances have significantly improved. This shows the importance of our proposed method for dealing with high dimensionality data under the imbalance problem.

Figure 4 shows each classification algorithms' performance for Dataset 2, before and after using PCA, SMOTE, and PCA+SMOTE.

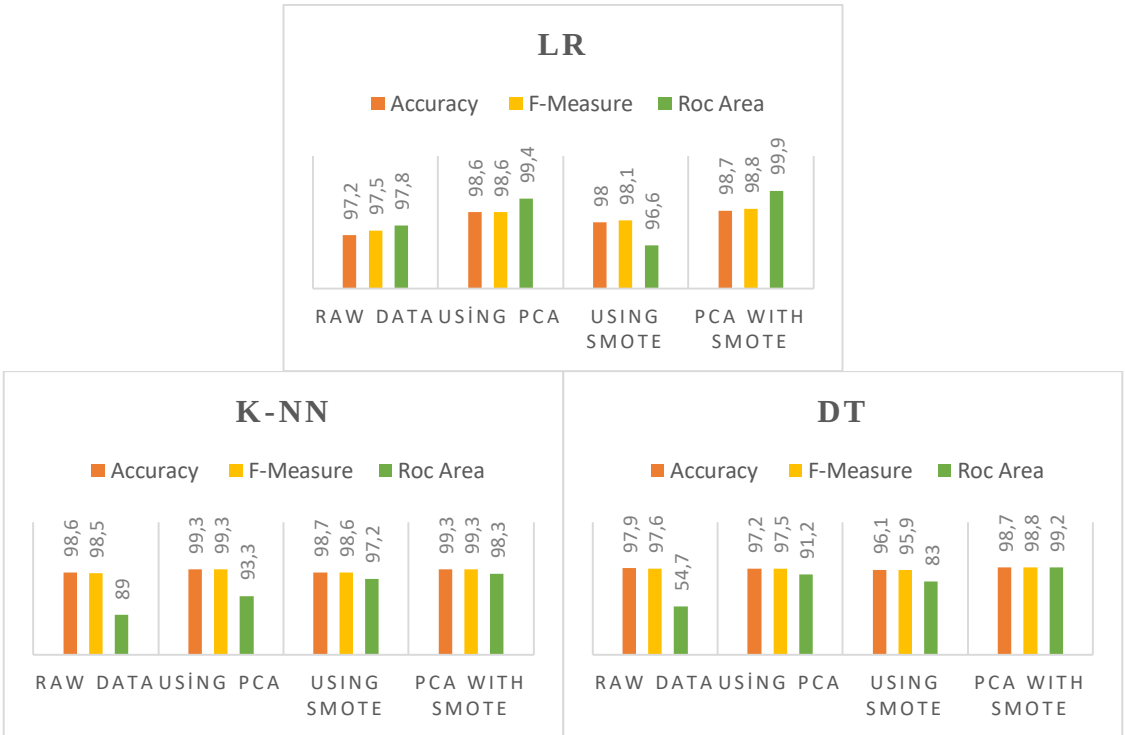

Figure 4. Performance of each classification algorithms for Dataset 2

In the Figure 4, For the LR algorithm, we can see that the Accuracy, F-measure and ROC are increased to 98.6%, 98.6% and 99.4%, respectively, in PCA method. When we used SMOTE method and the combination of PCA with SMOTE method, the LR algorithm has been improved. Using the PCA method, the K-NN algorithm has provided better results, but the DT algorithm has the worst results. While using the SMOTE method, the K-NN algorithm increases the classification results, but the results were decreased in the DT algorithm. The results in K-NN and DT have improved when the PCA's combined method with SMOTE is applied.

4. Conclusion and Recommendations

The issue was argued by decreasing the number of unnecessary features (using the PCA method of reducing dimensionality) and thus re-balancing data (using the SMOTE method). Hence, the health sector will benefit by developing a strategy for rapid and more accurate model identification to identify and address, along with efficient implementation [24].

Results on the three different classification algorithms for two imbalanced high-dimensional data showed that all classification algorithms have enhanced the classification performance of datasets using either PCA, SOMTE, or PCA+SMOTE methods. However, the preferred classification algorithm with the highest performance compared to others, varied from dataset to another. The first dataset's experimental results demonstrated that the best classification model was K-NN when normal classification, PCA, SMOTE, and PCA+SMOTE methods were used. However, in case of PCA, SMOTE and PCA+SMOTE methods, the results obtained from K-NN were not very good compared to other algorithms. On the other hand, when the PCA method was used, the DT algorithm's accuracy was improved, but K-NN and LR models did not show that improvement. Additionally, the other two classification algorithms' performances except the K-NN algorithm have been improved when using the SMOTE and PCA+SMOTE methods.

All algorithms showed improvement in SMOTE, PCA, and combined (PCA with SMOTE) methods in dataset 2. In this dataset, the best results for the raw data and the other three methods were obtained from the K-NN algorithm. When using PCA combined with SMOTE and SMOTE methods, all three algorithms showed significant improvement.

The experimental results from analyzed data indicated that when using PCA, SOMTE, or PCA+SMOTE methods, datasets classification performance is improved. However, when using these three methods, the K-NN model showed higher performance than the other two algorithms [4].

This study will contribute to the studies to be done with high dimensional unbalanced data in different fields. As a result, future research might focus on extending this approach to other real-world issues. Thus, besides an effective application in the field of health, it will be beneficial by developing an effective strategy in diagnosing and defining more accurate models in diagnosis and diagnosis.

Authors' Contributions

All authors contributed equally to the study.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The author declares that this study complies with Research and Publication Ethics

References

[1] Baran M. 2020. Makine Öğrenmesi Yöntemleriyle Çoklu Etiketli Verilerin Sınıflandırılması. Yüksek Lisans Tezi, Sivas Cumhuriyet Üniversitesi, Sosya Bilimler Enstitüsü, Sivas.

[2] Lorena A.C., Garcia L.P.F., Lehmann J., Souto M.C.P., Ho T.K. 2019. How Complex is Your Classification Problem?: A Survey on Measuring Classification Complexity. ACM Computing

Surveys, 52 (5): 1–34.
[3] Tahir M.A.U.H., Asghar S., Manzoor A., Noor M.A. 2019. A Classification Model for Class Imbalance Dataset Using Genetic Programming. IEEE Access, 7: 71013-71037.
[4] Mustafa N., Li J.P., Memon E.R.A., Omer M.Z. 2017. A Classification Model for Imbalanced Medical Data based on PCA and Farther Distance based Synthetic Minority Oversampling Technique. International Journal of Advanced Computer Science and Applications, 8 (1): 61-67.
[5] Kambhatla N., Leen, T.K. 1997. Dimension Reduction by Local Principal Component Analysis. Neural Computation, 9 (7): 1493-1516.
[6] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. 2009. The WEKA Data Mining Software: An Uptade. SIGKDD Explorations, 11 (1): 10-18.
[7] Sun Y., Wong A.K.C., Kamel M.S. 2009. Classification of Imbalanced Data: A Review. International Journal of Pattern Recognition and Artificial Intelligence, 23 (4): 687-719.
[8] Basgall M.J., Hasperué W., Naiouf M., Fernández A. 2018. SMOTE-BD: An Exact and Scalable Oversampling Method for Imbalanced Classification in Big Data. Journal of Computer Science & Technology, 18 (3): 203-209.
[9] Mohammed A.J., Hassan M.M., Kadir D.H. 2020. Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method. International Journal of Advanced Trends in Computer Science and Engineering, 9 (3): 3161-3172.
[10] Mythili M.S., Shanavas A.R.M. 2014. An Analysis of Students' Performance using Classification Algorithms. IOSR Journal of Computer Engineering, 16 (1): 63-69.
[11] Iyer A., Jeyalatha S., Sumbaly R. 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process, 5 (1): 1-14.
[12] Agrawal S., Agrawal J. 2015. Survey on Anomaly Detection using Data Mining Techniques. Procedia Computer Science, 60 (1): 708-713.
[13] Haghanikhameneh F., Shariat Panahy P.H., Khanahmadliravi N., Mousavi S.A. 2012. A Comparison Study between Data Mining Algorithms over Classification Techniques in Squid Dataset. International Journal of Artificial Intelligence, 9 (12): 59-66.
[14] Peng C.Y.J., Lee K.L., Ingersoll G.M. 2002. An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research, 96 (1): 3-14.
[15] Yıldız M., Bozdemir M.N., Kılıçaslan I., Atesçelik M., Gürbüz Ş., Mutlu B., Onur M.R., Gürger M. 2012. Elderly trauma: The two years experience of a University-affiliated Emergency Department. European Review for Medical and Pharmacological Sciences, 16 (SUPPL.1): 62-67.
[16] Samanthula B.K., Elmehdwi Y., Jiang W. 2015. K-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data. IEEE Transactions on Knowledge and Data Engineering, 27 (5): 1261-1273.
[17] Fix E., Hodges J.L. 1951. Discriminatory Analysis: Nonparametric Discrimination, consistency properties. Prepared at the University of California, Contract No, AF41, Texas. 43.
[18] Zhang Z. 2014. Too much covariates in a multivariable model may cause the problem of overfitting. Journal of Thoracic Disease, 6 (9) E196-E197.
[19] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J.O., Olakanmi O., Akinjobi J. 2017. Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology, 48 (3): 128-138.
[20] Mitchell T.M. 1999. Machine Learning and Data Mining. To Appear in Communications of the ACM, 42 (11): 1-13.
[21] Mohammed M., Khan M.B., Bashier E.B.M. 2017. Machine Learning Algorithms and Applications. Crc. Press, Bota Raton, 1-212.
[22] Prati R.C., Batista G.E., Monard M. 2009. Data mining with imbalanced class distributions: Concepts and methods. 4th Indian International Conference on Artificial Intelligence (IICAI-09), 16-18 December 2009, Tumkur India, 359-376.
[23] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16: 321-357.
[24] Naseriparsa M., Kashani M.M.R. 2013. Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset. International Journal of Computer Applications, 77 (3): 33-38.