# Empirical Study of Sampling Methods for Classification in Imbalanced Clinical Datasets

Asem Kasem[1], A. Ammar Ghaibeh[2(✉)], and Hiroki Moriguchi[2]

[1] School of Computing and Informatics, Universiti Teknologi Brunei, Gadong, Brunei Darussalam
asem.kasem@utb.edu.bn
[2] Medical Informatics Department, The University of Tokushima, Tokushima, Japan
ammargh@acm.org, h_moriguchi@ap6.mopera.ne.jp

**Abstract.** Many clinical data suffer from data imbalance in which we have large number of instances of one class and small number of instances of the other. This problem affects most machine learning algorithms especially decision trees. In this study, we investigated different undersampling and oversampling algorithms applied to multiple imbalanced clinical datasets. We evaluated the performance of decision tree classifiers built for each combination of dataset and sampling method. We reported our experiment results and found that the considered oversampling methods generally outperform undersampling ones using AUC performance measure.

**Keywords:** Clinical data mining · Imbalanced data · Undersampling · Oversampling · Decision tree C4.5

## 1 Introduction

There has been an increase of the use of data mining techniques in the different areas of medicine (bioinformatics, medical imaging, clinical informatics, and public health informatics) in the last decade. This is due to the impact data mining had on other domains, such as banking, marketing, and e-commerce, which gave high hopes for similar achievements in medicine, by extracting untapped knowledge contained in available medical data as well.

The aim of clinical data mining is to search for useful patterns and information within patients' data, and develop prediction models that can support clinical decision making [1, 2]. Data mining can be used to build predictive models in prognosis, diagnosis and treatment planning. Even when the data is collected for purposes other than directly diagnosing a disease or predicting treatment outcome, useful medical information can still be retrieved. Nakamura et al. [3] used data mining to predict the

First and second authors have contributed equally to produce this paper.


S. Phon-Amnuaisuk et al. (eds.), *Computational Intelligence in Information Systems*, Advances in Intelligent Systems and Computing 532, DOI 10.1007/978-3-319-48517-1_14

development of pressure ulcer in hospitals from patients' data that were originally collected for the purpose of calculating nursing costs. Decision making in the medical field is rather more sensitive than many other fields because of its direct relation to life and death consequences, and the well-being of patients. Therefore, a decision should be made with strong belief that is supported by thorough evaluation and clear explanation. This makes clinical data mining distinctive than other data mining uses in various ways. For example, it is widely common in clinical data mining to use white-box classifiers such as rule-based learners or decision trees because the resulting model is represented in a readable format. This enables the physicians to interpret the model output based on their medical knowledge, and increases their confidence when making their final decisions. While models built using black-box methods such as Artificial Neural Networks or Support Vector Machines and provide better results in terms of prediction accuracy, will be welcomed in many other fields, our experience showed that physicians often hesitate to accept these results due to the lack of model understandability, how the involved factors are related, and how to link that to their medical experience and knowledge. Although researchers have investigated the use of many different machine learning algorithms on clinical data, and reported interesting findings [1, 2], we believe that in practice, the ability for model introspection will actually limit us to only few of the many algorithms that are used in other fields and applications, even if this comes at the expense of prediction accuracy.

Another point to consider is that in many data mining applications, it is desirable to have a prediction model with high accuracy. In clinical data mining, however, it is important to distinguish between false positive errors and false negative ones. A false negative error has bigger impact than a false positive one because it can lead an unhealthy patient to miss a proper treatment, which might be fatal. On the other hand, a false positive error can be detected and corrected at a later stage by further investigations and tests.

Clinical datasets are usually highly heterogeneous where the data are usually collected from various sources such as images, laboratory tests, patient interviews, and physicians' observations and interpretations which leads to a poor mathematical characterization. In addition, many clinical datasets are noisy, incomplete, and suffer from the problem of data imbalance, in which the data has large number of patients (cases/instances) of one class (type/category), and a small number of patients of the other class.

In this study we consider using C4.5 decision tree, widely used in clinical data mining, with different sampling methods in order to identify best solutions for tackling the imbalanced data problem commonly faced in medical data mining.

The rest of this paper is organized as follows. In the next section, we explain decision tree classification models. We then discuss data imbalance problem and provide a description of common methods to overcome it in Sect. 3. Section 4 presents the clinical datasets considered in this study. Section 5 discusses the methodology we follow to conduct our experiments. Analyzing the results and reporting our findings is in Sect. 6. Finally, Sect. 7 concludes the paper and gives directions for future works.

## 2 Decision Trees

A decision tree model [4] is a data structure that is capable of representing knowledge in a humanly understandable way. It consists of a set of internal nodes, each representing test conditions on the values of one data attribute. The tree emerges from one common root node and ends with many leaf nodes, where each leaf represents a final classification decision.

Being able to understand how the built model is classifying the data and to interpret that into useful domain knowledge are the main reasons why decision trees are preferable over other methods like SVM or neural networks in clinical data mining [5]. A new data instance can be classified by starting from the root of the decision tree, and moving down its branches according to its attributes test results until a leaf node is reached. The class of the leaf node represent the predicted class of the instance. Attributes selected as a node test are usually determined using some splitting criteria. However, popular splitting criteria such as information gain ratio [4, 6] and Gini measure are skew sensitive.

## 3 Data Imbalance

Data imbalance is a problem that is very common in clinical data mining. A data set is considered imbalanced if the number of instances of one class is considerably smaller than the number of instances in the others. In clinical data the majority class is usually the negative class and the minority class is the positive class which is the class of our main interest. Multi-class problems might also suffer from data imbalance; however, it can be easily converted into many one-versus-others problem. Many learning algorithms tend to get overwhelmed by the large number of the majority class and ignore the minority class thus provide a high total accuracy, however, it also provides a high error rate on the minority class which is usually our concern. Assuming a 90 % imbalance ratio, a classifier that classify all instance as negative will achieve a 90 % accuracy while misclassifying all positive instances of the important class. Obviously this is not the desired result and some alternation is required to overcome this problem.

Japkowicz and Stephen [7], showed that different learning algorithms have different level of sensitivity to the data imbalance problem. They showed also that decision trees is the most sensitive classifier compared to Multilayer Perceptron and Support Vector Machines. In clinical data mining, decision trees are preferable because they provide an explanation of the classification decision.

### 3.1 Undersampling

Undersampling achieves data balance by removing instances from the majority class. Random undersampling method is the simplest form of undersampling in which the size of the majority class is reduced by removing instances randomly as its name indicates. Random undersampling is simple and easy to implement however, a main disadvantage of data undersampling methods is that there is a possibility that we lose

information contained in important majority class instances removed due to the undersampling process. A good informed-undersampling method reduces this possibility.

Informed undersampling reduces the size of the majority class in a controlled fashion in order to keep important instances from the majority class. Example of informed sampling are EasyEnsemble and BalanceCascade reported in [8]. Both methods use ensemble learning in order to explore the majority class space and select useful instances, however, ensemble learners models are usually difficult to explain and fall in the black-box learners zone.

J. Zhang and I. Mani [9] proposed four sampling methods called NearMiss-1, NearMiss-2, NearMiss-3, and Most-Distance that uses K-nearest neighbor in order to sample reduce the size of the majority class. The K-nearest neighbor of an instance is defined as the K elements whose distance between itself and the instance is the smallest. Here we provide a description of the four algorithms:

- *NearMiss-1* selects from the majority class the instances whose average distances to the three closest minority instances are the smallest. Thus the instances selected by NearMiss-1 are close to some of the minority class instances.
- *NearMiss-2* selects from the majority class the instances with the smallest average distance to the three farthest minority class. In other words, NearMiss-2 selects the majority instances close to all of the minority instances.
- *NearMiss-3* surrounds each instance form the minority class with $k$ instances from the majority class. It selects a predetermined number of the closest majority instances for each minority instance.
- *Most Distance* selects the instances from the majority class that have the largest average distance to the three closest instances from the minority class.

### 3.2 Oversampling

As its name indicates, oversampling works by sampling more data from the minority class. Random oversampling randomly selects a set of minority class $S_r$, duplicates its members, and appends them to the original minority class set. This will lead to an increase in the size of the minority class by the size of $S_r$ and a reduction in the original data imbalance distribution the process is repeated until the desired data balance reached. The problem with oversampling is that it may make the classifier susceptible to data overfitting because repeating the same instance causes the classifier to become more specific in covering these instances.

Another method of increasing the size of the minority class is synthetic sampling in which artificial data is synthesized from the original minority class. A powerful method that has shown good results in many applications is the synthetic minority oversampling technique (SMOTE) [10]. SMOTE uses feature space similarities between minority class instances in order to generate the synthesized artificial data. For each instance in the minority class in order to create a synthesized instance SMOTE randomly selects one of its K-nearest neighbor for some specified K, calculate the feature vector difference between the two instances then multiplies it by a random number in

the range [0, 1] and add the resulted vector to the original minority instance to generate the new artificial instance.

### 3.3 Model Evaluation

It is important to validate the model performance. Usually, accuracy is the evaluation metrics used to evaluate classification models. However, accuracy assumes similar cost for false positive and false negative errors. In clinical data mining, the cost of false positive is more expensive than the cost of false negative errors, and an evaluation method that reflects this fact is required.

Evaluation metrics are usually derived from the confusion matrix shown in Table 1.

From the confusion matrix, accuracy can be calculated as the ratio of correctly classified instances: Accuracy = (TP + TN) / ($P_c$ + $N_c$), and the classification error equals 1- Accuracy, i.e. Error = (FP + FN) / ($P_c$ + $N_c$).

**Table 1.** Confusion matrix

| | Positive Prediction: $P_p$ | Negative Prediction: $N_p$ |
|---|---|---|
| Positive Class: $P_c$ | TP: True Positive | FN: False Negative |
| Negative Class: $N_c$ | FP: False Positive | TN: True Negative |

Sensitivity and specificity can provide better metrics in the case of imbalanced datasets. Sensitivity, defined as TP / (TP + FN) = TP / $P_c$, measures the proportion of positive instances that are correctly classified.

On the other hand, specificity, defined as TN / (TN + FP) = TN / $N_c$, measures the proportion of negative instances that are correctly classified.

A good classifier should have high values for both sensitivity and specificity. In the case of imbalanced data, a classifier that classifies all instances as negative will have high accuracy, and high specificity, but zero sensitivity.

The Area Under Curve (AUC) [11] is widely used for measuring the performance in case of imbalanced data. AUC returns the area under Receiver Operating Characteristics Curve (ROC) that provides a visual representation of the performance in regards to the true positive rate (i.e. sensitivity) and false positive rate (i.e. 1-specificity). The visual presentation is useful for showing the tradeoffs between true positive and false positive error rates, however, it is difficult to use for calculation. The AUC provides a quantitative metric for ROC.

## 4 Experimental Datasets

Earlier experimental studies on learning from imbalanced data have been conducted. Reference [12] discussed the use of several sampling techniques versus different machine learners and performance metrics, and reported partial results of applying combinations of these choices on 35 datasets coming from a variety of application domains. In another study [13], the researchers investigated the class-imbalance

problem in medical datasets by considering different under-sampling and over-sampling techniques applied on one cardiovascular dataset. In this paper, we will investigate the effect of a group of undersampling and oversampling techniques applied on multiple clinical datasets, and under constraints suitable for data mining in the medical domain, where white-box learners and suitable metrics are of concern.

In our empirical study, we have considered 7 nonproprietary clinical datasets publically available in the following sources:

- UCI: the data repository of the Center for Machine Learning and Intelligent Systems in the University of California, Irvine, famously known as UCI Machine Learning Repository. (archive.ics.uci.edu/ml.)
- OML: an open collaborative machine learning platform (www.openml.org).

**Table 2.** Description of used clinical datasets

| Data set ID | Description | Source URL[a] |
|---|---|---|
| BRC | Breast Cancer dataset, from Institute of Oncology University Medical Center Ljubljana, Yugoslavia [14]. It was donated in 1988, and it is used to predict recurrent cancer events of patients. | Breast + Cancer |
| BCW | Breast Cancer Wisconsin dataset, donated from University of Wisconsin Hospitals in 1992 [15]. It is used to diagnose benign and malignant breast cancers. | Breast + Cancer + Wisconsin + %28Original%29 |
| DIB | Pima Indians Diabetes Database, donated in 1990, for predicting whether patients show signs of diabetes according to World Health Organization criteria. | Pima + Indians + Diabetes |
| SAH | South Africa Heart Disease dataset, taken from a larger dataset described by Rousseauw et al. in 1983. | www.openml.org/d/1498 |
| SPF | Heart dataset of cardiac Single Proton Emission Computed Tomography (SPECT) images, donated in 2001, where features are extracted from the images and used to predict cardiologists' diagnoses of normal and abnormal patients. | SPECTF + Heart |
| SPT | Same classification task as SPF, with binary extracted features to form the dataset. | SPECT + Heart |
| TYR | Thyroid Disease dataset, donated by the Garavan Institute in1987, to diagnose patients with thyroid disease. | Thyroid + Disease |

[a]Use archive.ics.uci.edu/ml/datasets/ before the value for UCI based datasets.

We have considered only datasets with binary classification problem. Table 2 lists these datasets with a brief description of each one, and an identifier to refer to later in our analysis.

The imbalance ratio, defined here as the percentage of minority class instances to majority class instances, varies from 9 % (highly imbalanced) to almost 54 % (only slightly imbalanced). The datasets have also diversity in the number of attributes, their types (continuous and categorical), and the number of instances.

Few datasets contain missing values in one or more of their attributes. In our study, we did not apply any method to fill in these values, and decided to work on complete data by removing the instances with missing values since they were only few. The TYR dataset was the only one having some attributes completely empty or redundant (these attributes were removed), and had rather big number of instances with missing values in some other attributes. The instances in the latter case have been removed, which we consider relatively acceptable given the total number of instances in this dataset. Table 3 summarizes these details.

## 5 Experiment Design

We have used RapidMiner (6.5) [16] to conduct our experiments. We have also used SMOTE implementation in Weka (3.16.13) software [17] to perform oversampling.

We have systematically applied each of the sampling methods, including "No Sampling", on each of the seven datasets. After performing data pre-processing, 10-folds cross-validation (stratified) was used in order to evaluate each method. In each fold, data balancing methods were applied to the training subset, while the test subset was left imbalanced. Figure 1 shows a snapshot of the cross-validation design in RapidMiner which limits the sampling application to the training set. AUC, sensitivity, and specificity were recorded for each sampling method.

**Table 3.** Datasets pre-processing and summaries

| Data set | # of [a] instances | # of [b] attributes | Attribute types | Instances removed | # of positive | # of negative | Imbalance ratio |
|---|---|---|---|---|---|---|---|
| BRC | 277 | 10 | all nominal | 9 | 81 | 196 | 0.41 |
| BCW | 683 | 10 | all numeric | 16 | 239 | 444 | 0.54 |
| DIB | 768 | 9 | all numeric | 0 | 268 | 500 | 0.54 |
| SAH | 462 | 10 | 8 numeric, 1 nominal | 0 | 160 | 302 | 0.53 |
| SPF | 267 | 45 | all numeric | 0 | 55 | 212 | 0.26 |
| SPT | 267 | 23 | all nominal | 0 | 55 | 212 | 0.26 |
| TYR | 2,643 | 23 | 6 numeric, 16 nominal | 1,129 | 212 | 2,431 | 0.09 |

[a]Final number of instances after removing instances with missing values.
[b]Number of attributes, including the class attribute, after pre-processing.

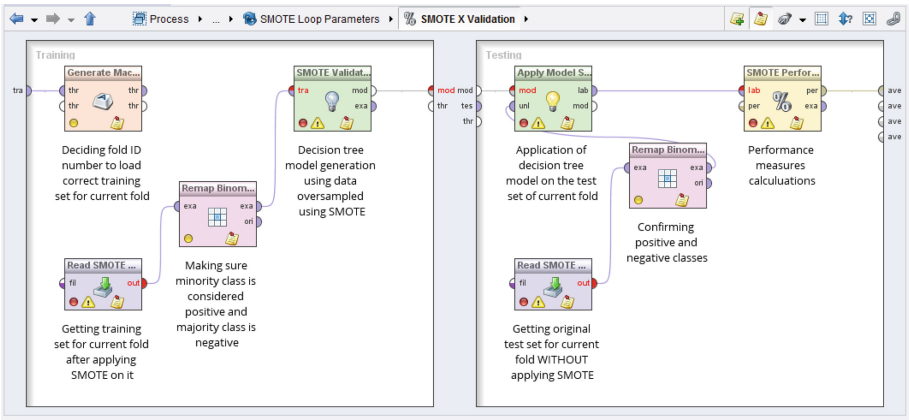


**Fig. 1.** Cross-validation process to evaluate SMOTE oversampling using RapidMiner.

The four sampling methods, NearMiss1, NearMiss2, NearMiss3, and Most Distance depend on calculating the distance between instances. In case of numeric attributes, Euclidean distance is used. However, when we have mixed types of attributes (numerical and categorical), Mixed-Euclidean distance is used, where for nominal attributes a distance of one is counted if corresponding values are not the same. Those algorithms are not part of RapidMiner components, and have been implemented by the authors.

For all sampling methods, we chose the parameters that rebalance the datasets to an almost equal ratio for both classes. As for the $k$ parameter (number of nearest neighbors) for SMOTE and NearMiss3 algorithms, a fixed value of 5 has been chosen.

## 6 Results and Analysis

The results of our experiments are shown in Tables 4, 5, 6 and 7. For each dataset, the area under curve AUC, sensitivity, and specificity of each of the methods used rounded to two decimal places are reported.

For the Breast Cancer (BRC) dataset, Table 4. (left) shows that Most Distance method scored the highest AUC, with corresponding 0.54 sensitivity and 0.81 specificity. It shows a good improvement in sensitivity over the results obtained on the original data (indicated by **No Sampling** method) with a relatively low reduction in specificity. Table 4. (right) shows the results for the BCW dataset, and the results for the remaining datasets are summarized in Tables 5, 6 and 7.

In Table 8, we have summerized the ranking counts for each method. For example, Random Undersampling method was ranked first in only one dataset, and similarly for second, third, and fourth ranks. It also ranked fifth in three datasets.

**Table 4.** Performance ranks and results on BRC (left) and BCW (right) datasets.

| # | Method | AUC | Sens. | Spec. |
|---|---|---|---|---|
| 1 | Most Distance | 0.69±0.11 | 0.54±0.16 | 0.81±0.06 |
| 2 | NearMiss1 | 0.68±0.09 | 0.53±0.15 | 0.74±0.10 |
| 3 | **No Sampling** | 0.65±0.11 | 0.32±0.19 | 0.92±0.05 |
| 4 | Rand. Under. | 0.65±0.09 | 0.40±0.20 | 0.74±0.18 |
| 5 | NearMiss3 | 0.64±0.11 | 0.30±0.15 | 0.91±0.06 |
| 6 | Rand. Over. | 0.63±0.08 | 0.42±0.11 | 0.83±0.16 |
| 7 | SMOTE | 0.63±0.10 | 0.40±0.18 | 0.84±0.07 |
| 8 | NearMiss2 | 0.62±0.09 | 0.69±0.11 | 0.46±0.11 |

| # | Method | AUC | Sens. | Spec. |
|---|---|---|---|---|
| 1 | Rand. Over. | 0.96±0.02 | 0.97±0.04 | 0.96±0.03 |
| 2 | SMOTE | 0.96±0.02 | 0.96±0.03 | 0.95±0.03 |
| 3 | **No Sampling** | 0.96±0.03 | 0.95±0.06 | 0.97±0.02 |
| 4 | NearMiss2 | 0.96±0.03 | 0.95±0.06 | 0.96±0.02 |
| 5 | Rand. Under. | 0.96±0.02 | 0.97±0.04 | 0.94±0.02 |
| 6 | NearMiss1 | 0.95±0.03 | 0.96±0.06 | 0.94±0.02 |
| 7 | Most Distance | 0.94±0.03 | 0.97±0.03 | 0.91±0.04 |
| 8 | NearMiss3 | 0.91±0.05 | 0.96±0.04 | 0.86±0.08 |

**Table 5.** Performance ranks and results on DIB (left) and SAH (right) datasets.

| # | Method | AUC | Sens. | Spec. |
|---|---|---|---|---|
| 1 | Rand. Over. | 0.71±0.07 | 0.49±0.28 | 0.80±0.21 |
| 2 | SMOTE | 0.71±0.07 | 0.71±0.25 | 0.61±0.23 |
| 3 | Most Distance | 0.66±0.05 | 0.90±0.06 | 0.43±0.06 |
| 4 | **No Sampling** | 0.66±0.07 | 0.28±0.08 | 0.95±0.05 |
| 5 | Rand. Under. | 0.66±0.08 | 0.42±0.26 | 0.83±0.19 |
| 6 | NearMiss1 | 0.62±0.05 | 0.38±0.07 | 0.90±0.04 |
| 7 | NearMiss3 | 0.60±0.05 | 0.24±0.08 | 0.96±0.02 |
| 8 | NearMiss2 | 0.50±0.00 | 0.53±0.07 | 0.50±0.09 |

| # | Method | AUC | Sens. | Spec. |
|---|---|---|---|---|
| 1 | Most Distance | 0.60±0.08 | 0.76±0.13 | 0.42±0.11 |
| 2 | Rand. Under. | 0.59±0.08 | 0.44±0.38 | 0.68±0.33 |
| 3 | SMOTE | 0.59±0.05 | 0.96±0.06 | 0.22±0.07 |
| 4 | **No Sampling** | 0.58±0.07 | 0.06±0.07 | 0.96±0.06 |
| 5 | Rand. Over. | 0.58±0.05 | 0.75±0.37 | 0.40±0.28 |
| 6 | NearMiss3 | 0.51±0.03 | 0.06±0.09 | 0.95±0.05 |
| 7 | NearMiss1 | 0.51±0.02 | 0.26±0.13 | 0.80±0.10 |
| 8 | NearMiss2 | 0.51±0.02 | 0.35±0.11 | 0.59±0.12 |

We can see from the table that oversampling methods have ranked first and second more often than the undersampling ones, with random oversampling ranked first more than SMOTE method. Among the undersampling methods NearMiss1 and NearMiss2 methods have often scored low ranks compared to other methods.

**Table 6.** Performance ranks and results on SPF (left) and SPT (right) datasets.

| # | Method | AUC | Sens. | Spec. |
|---|---|---|---|---|
| 1 | Rand. Over. | 0.75±0.13 | 0.78±0.22 | 0.71±0.14 |
| 2 | SMOTE | 0.73±0.14 | 0.74±0.21 | 0.72±0.11 |
| 3 | Rand. Under. | 0.71±0.12 | 0.79±0.23 | 0.64±0.15 |
| 4 | Most Distance | 0.69±0.09 | 0.98±0.05 | 0.44±0.14 |
| 5 | **No Sampling** | 0.66±0.13 | 0.19±0.17 | 0.91±0.07 |
| 6 | NearMiss2 | 0.65±0.09 | 0.73±0.20 | 0.54±0.17 |
| 7 | NearMiss1 | 0.56±0.11 | 0.49±0.26 | 0.54±0.10 |
| 8 | NearMiss3 | 0.50±0.00 | 0.22±0.17 | 0.76±0.11 |

| # | Method | AUC | Sens. | Spec. |
|---|---|---|---|---|
| 1 | Rand. Under. | 0.77±0.11 | 0.77±0.16 | 0.76±0.07 |
| 2 | Rand. Over. | 0.77±0.10 | 0.78±0.15 | 0.68±0.10 |
| 3 | **No Sampling** | 0.76±0.11 | 0.50±0.32 | 0.85±0.10 |
| 4 | NearMiss3 | 0.75±0.08 | 0.59±0.14 | 0.82±0.10 |
| 5 | SMOTE | 0.71±0.11 | 0.68±0.24 | 0.69±0.12 |
| 6 | Most Distance | 0.70±0.09 | 0.85±0.14 | 0.51±0.08 |
| 7 | NearMiss1 | 0.54±0.04 | 0.83±0.33 | 0.17±0.21 |
| 8 | NearMiss2 | 0.52±0.03 | 0.65±0.43 | 0.32±0.34 |

**Table 7.** Performance ranks and results on TYR dataset.

| # | Method | AUC | Sens. | Spec. |
|---|---|---|---|---|
| 1 | SMOTE | 0.97 ± 0.01 | 0.97 ± 0.03 | 0.94 ± 0.02 |
| 2 | Rand. over. | 0.96 ± 0.03 | 0.93 ± 0.05 | 0.98 ± 0.01 |
| 3 | NearMiss3 | 0.95 ± 0.02 | 0.91 ± 0.06 | 0.98 ± 0.01 |
| 4 | **No Sampling** | 0.95 ± 0.04 | 0.93 ± 0.06 | 0.98 ± 0.01 |
| 5 | Rand. Under. | 0.94 ± 0.03 | 0.93 ± 0.05 | 0.95 ± 0.02 |
| 6 | NearMiss2 | 0.93 ± 0.03 | 0.90 ± 0.07 | 0.96 ± 0.01 |
| 7 | NearMiss1 | 0.93 ± 0.03 | 0.90 ± 0.07 | 0.96 ± 0.01 |
| 8 | Most Distance | 0.60 ± 0.02 | 0.97 ± 0.04 | 0.22 ± 0.03 |

**Table 8.** Methods Rankings

| Sampling method | Rank count | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Random Undersampling | 1 | 1 | 1 | 1 | 3 | 0 | 0 | 0 |
| NearMiss1 | 0 | 1 | 0 | 0 | 0 | 2 | 4 | 0 |
| NearMiss2 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 4 |
| NearMiss3 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 2 |
| Most Distance | 2 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| Random Oversampling | 3 | 2 | 0 | 0 | 1 | 1 | 0 | 0 |
| SMOTE | 1 | 3 | 1 | 0 | 1 | 0 | 1 | 0 |
| No Sampling | 0 | 0 | 3 | 3 | 1 | 0 | 0 | 0 |

## 7 Conclusion

In this work, we have evaluated the performance of different sampling methods on clinical data classification problem using the C4.5 decision tree due to its wide usage in clinical data mining. The methods of random oversampling and undersampling, SMOTE oversampling, NearMiss1, NearMiss2, NearMiss3, and Most Distance undersampling methods were investigated. The results showed that from the AUC point of view, random oversampling and SMOTE methods were superior to the undersampling methods.

## References

1. Bellazzi, R., Zupan, B.: Predictive data mining in medicine: current issues and guidelines. Int. J. Med. Inf. **77**(2), 81–97 (2008)
2. Bellazzi, R., Ferrazzi, F., Sacchi, L.: Predictive data mining in clinical medicine: a focus on selected methods and applications. WIREs Data Mining Knowl. Discov. **1**, 416–430 (2011)
3. Nakamura, Y., Ghaibeh, A.A., Setoguchi, Y., Mitani, K., Abe, Y., Hashimoto, I., Moriguchi, H.: On-admission pressure ulcer prediction using the nursing needs score. JMIR Med. Inf. **3** (1), e8 (2015)
4. Ouinlan, J.R.: Induction of decision trees. J. Mach. Learn. **1**(1), 81–106 (1986)
5. Setoguchi, Y., Ghaibeh, A.A., Mitani, K., Abi, Y., Hasimoto, I., Moriguchi, H.: Predictability of pressure ulcers based on operation duration, transfer activity, and body mass index through the use of an alternating decision tree. J. Med. Investig. **63** (2016)
6. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
7. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. J. **6**, 429–449 (2002)
8. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B Cybern. **39**(2), 539–550 (2009)
9. Zhang, Z., Mani, I.: KNN approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of International Conference on Machine Learning (ICML 2003), Learning from Imbalanced Data Sets Workshop (2003)
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**(1), 321–357 (2002)
11. Bradley, A.P.: The use of area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. **30**(7), 1145–1159 (1997)
12. Hulse, J.V., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings of the 24th International Conference on Machine Learning (ICML 2007), Corvalis, USA, pp. 935–942 (2007)
13. Rahman, M.M., Davis, D.N.: Addressing the class imbalance problem in medical datasets. Int. J. Mach. Learn. Comput. **3**(2), 224–228 (2013)
14. Zwitter, M., Soklic, M.: Breast cancer dataset. Data obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia (1988)
15. Mangasarian, O.L., Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News **23**(5), 1–18 (1990)
16. RapidMiner Studio. http://www.rapidminer.com
17. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann (2011)