Proceedings of the Federated Conference on Computer Science and Information Systems pp. 203–211
DOI: 10.15439/2017F82
ISSN 2300-5963 ACSIS, Vol. 11

Preprocessing compensation techniques for improved classification of imbalanced medical datasets

Agnieszka Wosiak and Sylwia Karbowiak
Lodz University of Technology
Institute of Information Technology
ul. Wolczanska 215
90-924 Lodz, Poland
Email: agnieszka.wosiak@p.lodz.pl

***Abstract*—The paper describes the study on the problem of applying classification techniques in medical datasets with a class imbalance. The aim of the research is to identify factors that negatively affect classification results and propose actions that may be taken to improve the performance. To alleviate the impact of uneven and complex class distribution, methods of balancing the datasets are proposed and compared. The experiments were conducted on five datasets - three binary and two multiclass. They comprise several data preprocessing methods applied on data and the classification with different techniques. The study shows that for some datasets there exists a combination of a certain preprocessing method and a classification technique which outperforms other approaches. For datasets with complex distribution or too many features the ratio of correctly predicted labels may be low regardless what resampling method and classification technique has been applied.**

***Index Terms*—imbalanced datasets, class imbalance, medical data analysis, data preprocessing techniques**

I. Introduction

CLASSIFICATION is one of the core terms in machine learning. It refers to the process of prediction of class labels for unclassified, new instances basing on the knowledge drawn from historical, classified records [1]. It consists of application of the algorithm of classification technique on the already labeled data to build the classification model and attempt to discover the dependencies lying behind class labels. Afterwards, new instances are examined and assigned to the predicted groups.

Depending on the characteristics and the quality of gathered data it may be impossible to build a perfect model. In many real life cases, especially in medicine, accessing and measuring desired parameters is either costly, cannot be done precisely or at all. Collecting a representative number of samples from each class may be difficult due to above mentioned factors or due to naturally occurring disproportions. When one class is represented by much larger number of samples than the other, we refer to the class imbalance problem. It commonly arises in medical databases - large number of samples concerns patients with frequent observations while records describing special cases that are of particular interest may occur rarely.

Medical data analysis is important due to its meaning for medical decision making and diagnosis [2], [3]. Many studies have been conducted on the topic of classification techniques and how to improve their performance [4]–[6], specifically when it comes to the treatment of imbalanced datasets, but no universal, highly performing solution has been discovered yet.

Applying machine learning on unevenly distributed or incomplete datasets, in particular medical cases and resulting consequences, is discussed in the paper. Uneven class distribution is one of the problems that gains researchers' attention since late 90s [7]. In October 2005 dealing with non-static, imbalanced and cost-sensitive data was announced one of the top 10 challenging problems in data mining research by the International Journal of Information Technology and Decision Making [8].

The aim of the paper is to identify factors that affect classification results and propose actions that may be taken to improve the classification performance in terms of imbalanced datasets. Different combinations of preprocessing methods and classification techniques were used with regard to differences in datasets' characteristics: the number of target classes (two or more), the imbalance ratio, the number of features and the ratio of missing values. Even though classification problems have been studied extensively over the past few years, no universal solution has been discovered. Nowadays, there is still no perfect approach of classification as applied to imbalanced datasets and the paper constitutes an independent contribution to the relevant literature.

The rest of the paper is organized as follows. Section II describes the preprocessing techniques that may be applied to balance uneven distribution in datasets. Section III corresponds to the methods used in the experimental part of the paper and is followed by the description of medical data used in the research (Section IV). Section V is dedicated to the experiments conducted on sample data and the results. Finally, in Section VI, the concluding remarks are discussed.

II. Handling Imbalanced Datasets

The imbalanced class distribution may be defined by the ratio of the number of instances from minority class to those from majority class [9]. Such inequality may occur in many medical problems, where the number of patients diagnosed with rare illnesses, requiring special therapy or treatment is much smaller than the number of patients who do not need it. In certain domains, the datasets may be highly imbalanced with the imbalance ratio of, for example, 1:10000 [7].

Classification methods may fail when applied to an imbalanced dataset. Learning algorithms attempt to reduce global quantities such as the error rate and do not take the data distribution into consideration. As a result, samples from the dominant class are well-classified whereas samples from the minority class tend to be misclassified.

Weiss and Provost [10] after performing classification with decision tree in imbalanced two-class problems investigated the correlation between imbalance ratio and classification results. They found out that better results are obtained in a relatively balanced sets. However, the degree of class imbalance that starts to hinder the performance cannot be explicitly defined. 1:1 population ratio may not be always the optimal distribution to learn from.

The main approach of handling data imbalance problem is resampling in order to obtain more even class distribution. It allows classifiers to perform as in standard conditions. It is a flexible, independent of the classifier solution that usually improves classifier performance. Three main techniques of datasets' balancing are described in Sections II-A – II-C.

A. Undersampling

Undersampling involves a removal of some examples from the majority class. Non-random selection of sample removal is called a focused undersampling and may refer to the samples of the majority class lying further away [11]. Two non-random examples of informed undersampling that proved to give good results are EasyEnsemble and BalanceCascade algorithms [7], [12]. Both of them intend to overcome information loss introduced in the traditional random undersampling method.

One of more interesting approaches that was applied in [14] is Neighbourhood Cleaning Rule (NCR). Given a sample in a training set, three nearest neighbors are found. If all neighbors belong to minority class while a sample belongs to majority class - the sample is removed. In the contrary case - when a sample belongs to the minority class and its three nearest neighbors to opposite class - all three neighbors are removed [23]. In other words NCR is an informed undersampling technique where majority class samples are removed only when they closely surround or are surrounded by minority class samples.

B. Oversampling

Oversampling consists of generating new examples and adding them to the original dataset. Similarly to undersampling, two approaches can be distinguished: random and focused oversampling. Random oversampling refers to simple replication of existing samples. Focused oversampling means oversampling only those minority examples that occur on the boundary between the minority and majority classes.

The main advantage of oversampling is no loss of information from original dataset. On the other hand, it increases dataset size and thus computational cost [20] and may result in overfitting due to too many *tied* instances [7]. Random undersampling carries a risk of missing potentially important data, however Drummond and Holte [21] show that random under-sampling yields better minority prediction than random over-sampling.

Garcia et al. [22] applied four resampling algorithms and eight different classifiers on 17 real datasets. Authors' experiment showed that oversampling the minority class outperforms undersampling the majority class when datasets are strongly imbalanced and there are not significant differences for data with a low imbalance. Results also indicated that the classifier had a very poor influence on the effectiveness of the resampling strategies.

A variation of oversampling called Synthetic Minority Oversampling Technique (SMOTE) was proposed in 2002 by N.Chawla et al. [24] which produces synthetic examples. New minority class examples are created along the line segments between each positive class object and any of the k-nearest neighbors.

SMOTE shows that a combination of oversampling the minority class and undersampling the majority class can achieve better classifier performance than only undersampling the majority class. It has proven good efficiency in many works but a problem may appear when a dataset is not only imbalanced but also has a complex distribution. In such a case synthetic samples generation may lead to the overlapping between classes.

C. Hybrid approach

Hybrid approach is a combination of over- and undersampling [24], eliminating some of the examples before or after resampling, in order to reduce overfitting. It allows to balance the dataset and keep the trade-off between decreasing majority class size and replication of minority class samples. Common approach is a combination of random undersampling with SMOTE.

D. Multiple imbalanced class problems

Datasets with more than two classes imply an additional difficulty for classification algorithms. When multiple labels are present, solutions proposed for binary-class problems may not be directly applicable, or may achieve a lower performance than expected. For example, solutions at data level suffer from the increased search space, and solutions at algorithm level become more complex, as the learning algorithm must consider several small classes [23].

Fernandez and Lopez [23] presented binarization schemes in order to apply standard approaches to solve two-class imbalanced problems as well as several procedures which have been designed for the scenario of imbalanced datasets

with multiple classes. They proposed to transform the original problem into binary subproblems.

Class binarization techniques make it possible to apply the standard classification solutions. Two best known approaches to transform a multiple class classification problem into a set of binary problems are distinguished.

a) One-versus-one (OVO): The approach trains a classifier for each possible pair of classes, ignoring the examples that do not belong to the related classes. When classifying instances, a query is submitted to all binary models, and the predictions of these models are combined into an overall classification. For those algorithms that do not have an associated certainty degree for each class, the most common way to generate the class label is to represent the output of each binary classifier in a code matrix.

b) One-versus-all (OVA): The approach builds a single classifier for each of the classes of the problem, considering the examples of the current class to be positives and the remaining instances negatives. An instance will be assigned to the majority class, or randomly among the majority classes if they have the same amount of examples.

E. Complex distribution

Additionally to class imbalance two other major factors with regard to class distribution can be distinguished: class overlapping and areas with small disjuncts and noise.

The serious problem that complicates learning of the minority class is a difficulty in separation of two classes. When in some feature space overlapping patterns are present, it is hard to determine rules for separating one class from another. Such a feature may become redundant to help recognize decision boundaries between classes.

Often standard classifiers that tend to maximize accuracy in classification fail while encountering the problem of overlapping, since they classify the overlapping region as belonging to the majority class and assume the minority class is noise [25], [26].

Another issue concerning class distribution is when a class consists of several sub-clusters of different amount of examples, referred to as small disjuncts. Many current approaches to class imbalance mostly aim to solve the between-class imbalance problem and disregard the uneven distribution within the class [27].

III. Methodology

The proposed methodology of indicating the best pairwise combination of the preprocessing technique of datasets' balancing and the classification method consists of three steps:

1) applying classification methods on the original dataset without preprocessing (NOP),
2) performing preprocessing on datasets,
3) carrying out classification on datasets modified in the previous step,
4) comparing results of classifications.

The datasets were modified with the following methods:

- random undersampling (RU),
- SMOTE (SM) as a variation of oversampling,
- hybrid approach by SMOTE and random undersampling (SM-RU).

Four classification techniques were applied on original and preprocessed datasets:

- decision tree (DT),
- Naïve Bayes (NB),
- k-nearest neighbors with k=3 (3NN) and k=5 (5NN) neighbors,
- support vector machine (SVM).

Due to the fact that the presented approach aims at supporting medical diagnosis, there were chosen simple, comprehensible algorithms, as physicians should understand the tools they use.

For kNN and SVM all string or category features were normalized and mapped to numerical values where necessary. The information for the value mapping was taken from the dictionaries built in a Predictive Model Markup Language files (PMML) and integrated with the experimental environment [13].

Additionally for the sets with incomplete data, the influence of substitution of missing values by mean values was examined (-SUB suffix).

Random undersampling was performed by random row removal from all classes until each class had the same number of samples as in the least numerous minority class. SMOTE used 5 nearest neighbors, which was also consistent with results described in [14].

The approach with random undersampling and SMOTE replicated minority data by 5 times and randomly removed rows from majority class to reach equal number of rows for every class.

Gini index was applied in decision tree classification to evaluate scores, 3 and 5 neighbors were used for kNN classifier and radial basis function (RBF) kernel was employed in SVM.

Experiments for each combination were repeated 10 times. Each original dataset was divided into test and validation sets in proportion 9:1 with 10-fold cross-validation, which is widely accepted in data mining and machine learning community and serves as a standard procedure of validation [15]–[17].

Accuracy and sensitivity were chosen as evaluation metrics. Sensitivity tells how good the technique is in determining the exact class label while accuracy gives an overall ratio of correct predictions to all predictions made. All values presented in the tables are mean values from the scores obtained in 10 runs. Additionally sensitivity score is a mean value from predictions of all classes: minority and majority ones.

IV. Data Description

To demonstrate the problems encountered while dealing with medical data, test cases with different characteristics have been chosen. All of them are public datasets dedicated to researchers for machine learning tasks and medical diagnosis improvement:

TABLE I: Characteristics of considered medical datasets

Dataset name	Number of samples	Class labels ratio	Number of attributes	Missing values
Hepatitis	155	123 : 32	19	5.7%
Lung Cancer	96	86 : 10	7129	0.0%
Hypotyroid	3163	3012 : 151	30	3.2%
Thyroid	7200	6666 : 368 : 166	21	0.0%
Lung SCC Cancer	494	467 : 14 : 13	71	8.6%

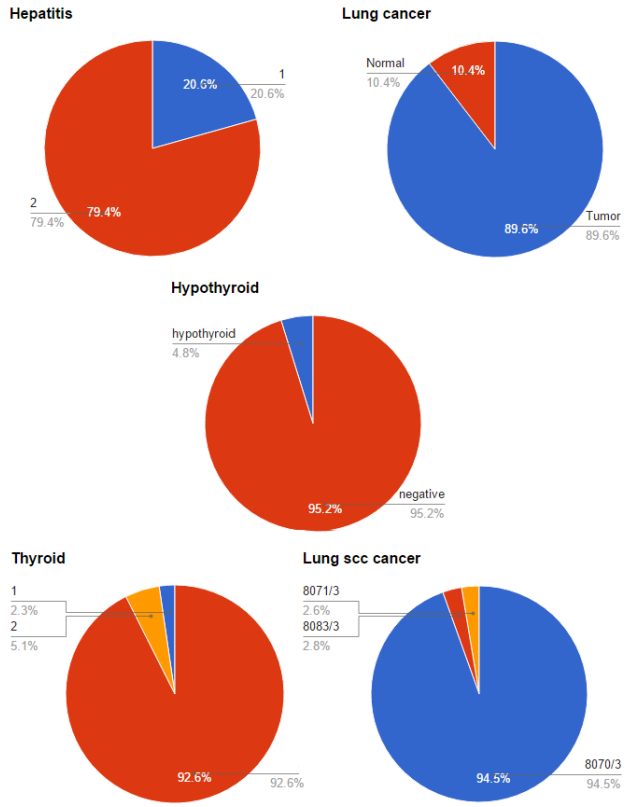

Fig. 1: Pie charts for class imbalance

- Hepatitis [28],
- Lung Cancer [29],
- Hypothyroid Disease [30],
- Thyroid Disease [31], and
- Lung Squamous Cell Carcinoma [32].

Table I presents a brief characteristics of the datasets. The class imbalance is presented in graphical form in the Figure 1. Further details are discussed separately in the subsequent paragraphs.

Hepatitis is a dataset with two classes where imbalance ratio is equal to 0.26. The class attribute determines whether patient is dead (32) or alive (123). All other attributes are numerical and represent age, sex and other indicators' values gathered by medical scientists. There are missing values - only one column misses more than 43% of values and others up to 19%.

Lung cancer is also a two class problem with imbalance ratio 0.12. It refers to lung cancer diagnosis. Minority class consists of 10 non-neoplastic (normal) lung samples and majority of 86 primary lung adenocarcinomas (tumor) samples. There are no missing values and each sample is described by 7129 genes (numeric attributes).

Hypothyroid is a two class problem with strong imbalance ratio, 0.05. Majority class holds attribute 'negative' while minority is diagnosed as 'hypothyroid'. The dataset has a relatively small number of missing values, but one of the columns with more than 90% of missing values was removed in preliminary data preparation.

Thyroid is a dataset with three classes; the most numerous class has over 18 times more samples than the first minority class and over 40 times more than the other minority class. The dataset is relatively big, there are not many attributes and no missing values.

Lung scc cancer dataset refers to Lung Squamous Cell Carcinoma cancer type. It contains samples described by numerical and nominal attributes and is characterized by high ratio of missing values. Classification in this sets is done by assigning International Classification of Diseases for Oncology, Third Edition ICO-3 Histology Code. The problem has one majority class (Squamous cell carcinoma) and two minority classes: Basaloid squamous cell carcinoma and Keratinizing squamous cell carcinoma.

Prior to proper data processing several rows from original Lung scc cancer dataset were removed due to their belonging to extremely rare class which will not be considered. Also, attributes that missed over 70% of values, carried identifiers, non-relevant information or the same value for all samples were filtered out (more than 20 columns in total). The process of excluding less relevant attributes in terms of further classification called a feature selection is gaining on popularity and was discussed in [35].

V. Results and Discussion

The purpose of experiments was to find how the preprocessing compensation methods improve classification of imbalanced medical datasets. The experiments were conducted according to the methodology introduced in Section III on public datasets described in Section IV.

The experiments were performed with use of The Konstanz Information Miner environment (KNIME), Version 3 [18], [19].

A. Experimental Results for Hepatitis Dataset

Sensitivity mean values for hepatitis dataset are presented in the Table II. The best score was obtained for Naïve Bayes classifier with a hybrid approach: random undersampling and SMOTE and kNN with 5 neighbors combined with random undersampling. Support Vector Machine and Naïve Bayes showed also a high performance when trained on datasets with reduced, equal number of samples for each of the classes. Decision tree algorithm was the worst in this classification no matter which preprocessing method was applied. It can be pointed out that substitution of missing values improved

TABLE II: Hepatitis sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.6183	0.7874	0.6708	0.7183	0.5506
RU	0.6704	0.7860	**0.7776**	**0.8045**	**0.7824**
SM	0.6436	0.7897	0.7520	0.7708	0.5000
SM_RU	0.6607	**0.8025**	0.6712	0.7287	0.5000
NOP_SUB	0.6305	0.6061	0.7672	0.7000	0.7568
RU_SUB	**0.7197**	0.6531	0.7546	0.7460	**0.7835**
SM_SUB	0.6692	0.6769	0.7533	0.7391	0.5000
SM_RU_SUB	**0.7147**	0.6110	0.7486	0.7398	0.5000

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling, SUB - substitution of missing values

TABLE III: Hepatitis accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.7574	**0.8277**	0.8225	0.8500	0.8288
RU	0.6897	0.7723	0.7988	0.8388	0.8225
SM	0.7426	0.8148	0.8338	0.8550	0.8375
SM_RU	0.7348	**0.8277**	0.8388	**0.8675**	0.8375
NOP_SUB	**0.7768**	0.7968	**0.8500**	0.8350	**0.8575**
RU_SUB	0.7497	0.5871	0.7500	0.7563	0.7775
SM_SUB	0.7594	0.5606	0.7425	0.7188	0.8375
SM_RU_SUB	0.7381	0.8045	0.7450	0.7200	0.8375

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling, SUB - substitution of missing values

its performance when random undersampling and hybrid sampling were applied on the dataset.

The accuracy scores (Table III) reach highest values for 3NN (with missing values substitution), 5NN and SVM - no resampling for all of them. Naïve Bayes' best accuracy is worse than for mentioned classifiers but significantly better than for weakly performing decision tree.

In the Figure 2, the highest accuracy scores obtained in the experiment are compared with accuracy scores in cases where sensitivity for given classifier was highest. Only Naïve Bayes with hybrid approach reaches highest sensitivity with highest accuracy.

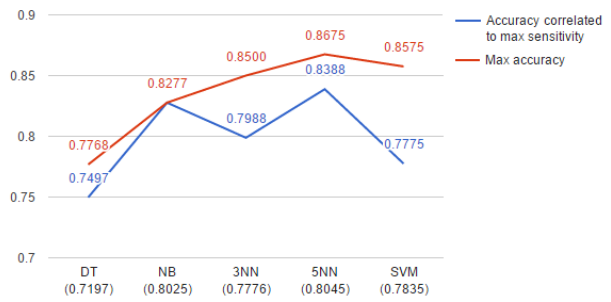

Fig. 2: Highest accuracy score vs. accuracy correlated with highest sensitivity score for hepatitis dataset

TABLE IV: Lung cancer sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	**0.9492**	0.5000	**0.9792**	0.8892	0.5000
RU	0.9357	0.5000	0.8703	0.8986	**0.6350**
SM	0.9286	0.5000	0.9442	**0.9407**	0.5000
SM_RU	0.9357	0.5000	0.9488	0.9343	0.5000

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling

TABLE V: Lung cancer accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	**0.9802**	**0.8958**	**0.9865**	**0.9677**	0.8960
RU	0.8927	0.1042	0.8073	0.8500	**0.9240**
SM	0.9719	0.1042	0.9000	0.8938	0.8958
SM_RU	0.9750	**0.8958**	0.9083	0.8823	0.8958

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling

B. Experimental Results for Lung Cancer Dataset

For lung cancer dataset sensitivity scores (Table IV) differed a lot across classifiers. The best result was achieved for kNN method. With 3 neighbors and no data preprocessing 9 out of 10 runs gave correct classification for whole minority class. Similar results were attained for 5NN classifier and they were only slightly worse than decision tree and hybrid over- and undersampling. SVM performed poorly but one better score was obtained when dataset was reduced. Naïve Bayes and support vector machine with other types of resampling were not capable to build any model correctly predicting the minority class labels.

This set has all records complete so no tests were made for classification with missing values substitution.

The accuracy scores in Table V were not correlated with sensitivity measure. In general, best values were obtained for classification in not preprocessed datasets with small exceptions for decision tree and SVM.

The comparison of the highest accuracies and the accuracies where the sensitivity was the highest is shown in the Figure 3. Almost all classifiers, except for 5NN, resulted in a high sensitivity and accuracy at the same time - a decision tree and 3NN with no preprocessing, SVM with random undersampling.

C. Experimental Results for Hypothyroid Disease Dataset

The results for hypothyroid dataset (Table VI) are similar for all classifiers but SVM. A decision tree performed well even if no preprocessing was applied. Other techniques attained the best results when data was either undersampled or also beforehand oversampled. SMOTE also improved classification correctness significantly when comparing to no preprocessing at all. It may be observed that substitution of missing values slightly improved the sensitivity for kNN and SVM.

For all the classifiers without exceptions the accuracy was the best in case of an original dataset and when missing values

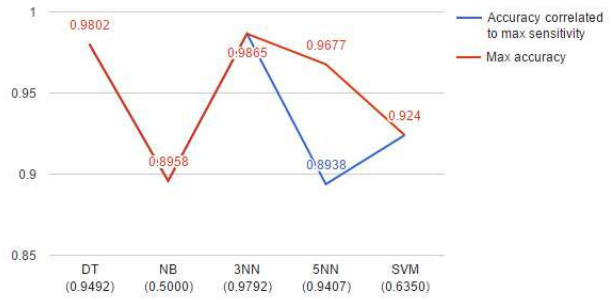

Fig. 3: Highest accuracy score vs. accuracy correlated with highest sensitivity score for hepatitis dataset

TABLE VI: Hypothyroid Disease sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	**0.9439**	0.8836	0.9126	0.8935	0.5934
RU	0.9295	**0.9490**	**0.9499**	0.9448	0.8532
SM	0.9216	0.9161	0.9442	0.9432	0.5000
SM_RU	**0.9386**	0.9272	**0.9474**	0.9441	0.5000
NOP_SUB	**0.9427**	0.8346	0.9017	0.8886	0.6206
RU_SUB	0.9152	0.8448	**0.9490**	**0.9495**	**0.8998**
SM_SUB	0.9129	0.8468	0.9426	0.9437	0.5000
SM_RU_SUB	0.9164	0.8427	**0.9523**	**0.9513**	0.5000

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling, SUB - substitution of missing values

were substituted by mean values (Table VII).

The comparison of the highest accuracies and the accuracies where sensitivity was the highest presented in the Figure 4 proves that decision tree without preprocessing is the most sensitive to the minority class and gives the most accurate predictions for both classes. For other classifiers where resampling was applied on a training dataset, the accuracy scores are slightly worse than the highest scores obtained.

D. Experimental Results for Thyroid Disease Dataset

The results for multi-class Thyroid disease data classification showed in the Table VIII vary across the methods applied. The best sensitivity scores were observed for a decision tree with random undersampling alone and when combined with

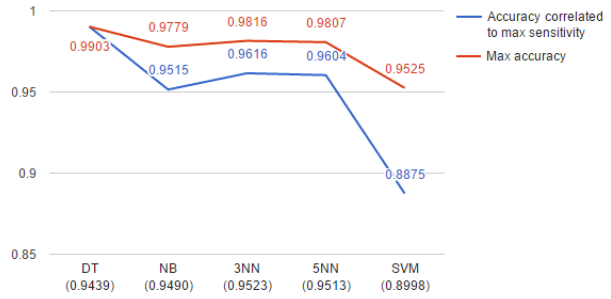

Fig. 4: Highest accuracy score vs. accuracy correlated with highest sensitivity score for hypothyroid dataset

TABLE VII: Hypothyroid Disease accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	**0.9903**	**0.9779**	**0.9806**	**0.9807**	**0.9497**
RU	0.9065	0.9515	0.9419	0.9402	0.8034
SM	0.9321	0.9708	0.9736	0.9711	0.9390
SM_RU	0.9370	0.9704	0.9625	0.9612	0.9390
NOP_SUB	**0.9862**	**0.9695**	**0.9816**	**0.9801**	**0.9525**
RU_SUB	0.8745	0.9368	0.9452	0.9461	0.8875
SM_SUB	0.9013	0.9646	0.9707	0.9670	0.9390
SM_RU_SUB	0.8929	0.9622	0.9616	0.9604	0.9390

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling, SUB - substitution of missing values

TABLE VIII: Thyroid Disease sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.9862	0.7051	0.5667	0.5403	0.4484
RU	**0.9926**	**0.8319**	0.6695	0.6745	**0.6320**
SM	0.9911	0.7826	0.6921	0.6994	0.3333
SM_RU	**0.9962**	0.8046	**0.7127**	**0.7056**	0.3333

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling

SMOTE. Next score was obtained by Naïve Bayes and RU, then 3NN, 5NN with the hybrid approach and finally again poorly performing SVM with random undersampling.

All the best accuracy values (table IX) were observed for all classifiers when applied on the original datasets.

The decision tree without preprocessing offers a perfect trade-off between maximum sensitivity and accuracy. In case of other classifiers improvement in sensitivity score causes a decrease of the accuracy (figure 5).

E. Experimental Results for Lung Squamous Cell Carcinoma Dataset

The Lung scc cancer is an experimental dataset with two minority classes. Only a decision tree reached outstanding sensitivity score when applied on an undersampled dataset (Table X). Naïve Bayes performed best combined with hybrid resampling approach, but the accuracy was still very low.

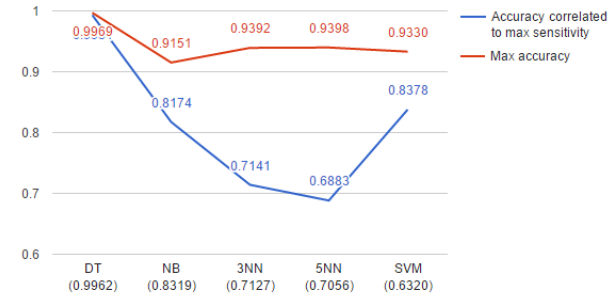

Fig. 5: Highest accuracy score vs. accuracy correlated with highest sensitivity score for thyroid dataset

TABLE IX: Thyroid Disease accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
	DT	NB	3NN	5NN	SVM
NOP	**0.9969**	**0.9151**	**0.9392**	**0.9398**	**0.9330**
RU	0.9805	0.8174	0.6134	0.6058	0.8378
SM	**0.9969**	0.8942	0.8662	0.8413	0.8732
SM_RU	0.9931	0.8821	0.7141	0.6883	0.9258

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling

TABLE X: Lung Squamous Cell Carcinoma sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.3350	0.3690	0.3333	0.3333	-
RU	**0.5493**	0.3405	0.2227	0.2581	-
SM	0.3727	0.3558	0.3018	0.2956	-
SM_RU	0.4282	**0.3851**	0.3333	0.3333	-
NOP_SUB	0.3457	0.3314	0.3333	0.3333	0.3333
RU_SUB	0.3891	0.3445	0.1935	0.1872	0.1614
SM_SUB	0.3979	0.3664	0.2406	0.3541	0.3333
SM_RU_SUB	0.3622	0.3644	0.2670	0.2420	0.3333

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling, SUB - substitution of missing values

For dataset with missing values it was not possible to find a hyperplane for the Support Vector Machine in a finite time, thus no sensitivity scores are presented in this section.

The accuracy for this dataset (table XI) are again mainly the best for no preprocessing.

For the lung scc cancer dataset the differences in the maximum accuracy and the accuracy when the sensitivity was the highest are presented in figure 6. Only a decision tree and Naïve Bayes were shown since other classifiers did not provide satisfactory sensitivity outcomes. Especially in the case of a decision tree the improvement in a sensitivity score cost a double drop in the accuracy score.

F. Discussion

In the conducted experiments five datasets were examined. The goal was to find out which preprocessing method and a

Fig. 6: Highest accuracy score vs. accuracy correlated with highest sensitivity score for lung scc cancer dataset

TABLE XI: Lung Squamous Cell Carcinoma accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	**0.9287**	0.9047	**0.9741**	**0.9741**	-
RU	0.4663	0.1636	0.6509	0.7543	-
SM	0.8830	0.8755	0.8819	0.8638	-
SM_RU	0.7621	0.6759	**0.9741**	**0.9741**	-
NOP_SUB	**0.9245**	**0.9399**	**0.9741**	**0.9741**	**0.9741**
RU_SUB	0.3802	0.1899	0.4698	0.4991	0.2802
SM_SUB	0.7812	0.2530	0.6552	0.6043	**0.9741**
SM_RU_SUB	0.6518	0.2474	0.6845	0.5638	**0.9741**

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM_RU - SMOTE and random undersampling, SUB - substitution of missing values

classification technique performs best under given conditions. All datasets represented class imbalance problem with different level of class labels distribution. There were binary and multiclass problems, with few or many samples and narrow or vast feature space. The aim was also to demonstrate how different characteristics influence performance of various data treatment methods - resampling and missing values imputation - and certain classification techniques.

Mean sensitivity and accuracy scores were given for each test on the combination of a resampling method and a learning algorithm. As already mentioned, accuracy may be not truly informative when assessing classifier's ability to identify minority samples. The correctly predicted labels mostly belong to majority class while minority class cases are frequently misclassified. Therefore a sensitivity score is more relevant as it indicates how good the predictions were within each class label. The classifiers with a high sensitivity, yet not the highest accuracy, are better in the identification of minority class samples. Consequently, the results of the experimental studies will be ranked by the sensitivity scores and accuracy will be considered as less significant.

For imbalanced two class Hepatitis dataset the best performing classification technique in terms of general accuracy and sensitivity to minority class samples was k-nearest neighbors. The best sensitivity scores were reached when combined with random undersampling. Naïve Bayes with hybrid preprocessing - random undersampling and SMOTE gave similar results to kNN. Most of other classifiers performed well when combined with random undersampling. Additionally, less efficient SVM and decision tree were more sensitive when missing values were substituted by mean values. As more than 5% of values were missing, mean value imputation usually improved the performance of classifiers. All kinds of classifiers trained on resampled datasets were more sensitive than without data preprocessing. The sensitivity and accuracy rates at the level of 70-85% suggest that learning algorithms cannot be considered as a truly reliable solution for the problem of classifying new instances.

The lung cancer dataset was also a two class problem. The characteristics of dataset revealed no missing values, high imbalance ratio and small sample size. Each instance was char-

TABLE XII: Compilation of correctly predicted labels for minority class (True Positives) and majority class (True Negatives)

	Actual number of samples	Predicted with NB NOP	Predicted with NB RU	Change
True Positive	151	117	142	16.56%
True Negative	3012	2977	2869	-3.59%

acterized by 7129 attributes. High dimensionality appeared to be a problem for Naïve Bayes and Support Vector Machine that were not able to create a proper probability model or decision surface with so many parameters in a reasonable classification time. Feature selection would probably help to decrease the dimensionality and improve their performance. kNN with k=3 performed the best taking into account both accuracy and sensitivity when data was not processed. It means that data is well structured and unlabeled samples are most often close to other samples of their actual class. For other classifiers different preprocessing methods significantly improved their ability to recognize minority samples without a rapid decrease of the accuracy.

Hypothyroid is the last of examined binary class problems. The sensitivity scores for all classifiers excluding the Support Vector Machine were similar and no best performing combination can be indicated. The highest sensitivity score was attributed to kNN with 3 and 5 neighbors and hybrid resampling. The number of rows affected by missing vales is lower than in case of the hepatitis dataset so a value imputation did not improve the sensitivity significantly. For all classifiers resampling improved sensitivity but accuracy scores remained at the highest level even when no preprocessing was applied. In order to better predict the samples from the minority class, a trade-off between improving the sensitivity and at the same time worsening the overall performance should be accepted. As an example, differences in correct labels predictions for last fold in final iteration of Naïve Bayes trained on a dataset with random undersampling (NB RU) versus trained on original dataset (NB NOP) were compiled in Table XII. It may be observed that after training on the dataset balanced with random undersampling, the classifier identified 1/6 more of minority samples and misclassified less than 4% of actual majority class instances.

Thyroid disease is a three-class problem. The best results were attained for a decision tree algorithm, no matter which preprocessing method was applied. It was due to the precisely defined split conditions and well separated minority class from majority one. Random undersampling with or without SMOTE improved the sensitivity scores for all classifiers tested, while accuracy remained best for datasets without preprocessing. Taking both metrics into account, decision tree with SMOTE reached best results for the problem.

The Lung scc cancer dataset has two minority classes and the third class significantly larger than two others. It could be observed that scores for any classification technique and preprocessing method performed worse in that case than in the previous scenarios. On average, a half of instances were classified correctly by a decision tree algorithm combined with random undersampling, which is even worse by two times when compared with algorithm applied on not balanced, original dataset. This is an extremely difficult classification problem since the dataset is highly imbalanced - each of positive class instances constitute less then 3% of number of negative instances, there are three class labels and a ratio of missing values is relatively high. No combination of preprocessing method and classification technique can be considered reliable while classifying a new instance. It could be stated the balancing did not succeed in terms of highly uneven distribution of instances between separate classes.

VI. Conclusions

Real-life medical datasets are often imbalanced, sparse and high-dimensional. Class imbalance is one of the key problems and it imposes additional difficulties on learning from data.

The point at issue is to what degree should one balance the original dataset or what kind of assumptions will make learning algorithms perform better than when considering the original distribution. The answer is open since this field still lacks a uniform benchmark platform and standardized performance assessments. Although there are many publicly available datasets, a very limited number concerns imbalanced class problems. Data sharing is not common and research groups are required to collect and prepare their own datasets [7]. There is still not much of theoretical understanding on the principles of this problem. Many algorithms that were proposed over years are able to improve classification accuracy over certain benchmarks but will fail over the others.

In the paper several classification techniques and data preprocessing methods were investigated. They were applied on datasets with various characteristics to distinguish factors and conditions that make a learning algorithm perform better. The application of resampling methods for imbalanced datasets enabled attaining higher results in terms of accuracy and sensitivity. The hybrid approach built by the combination of random removal of majority class samples and Synthetic Minority Oversampling Technique overcome single preprocessing techniques.

The paper considered simple, comprehensible algorithms that can be well understood by medical staff. However, in recent days an evolution from traditional learning algorithms towards neural networks and artificial intelligence solutions is observed [33]. Such methods may appear efficient but inability to identify the rules that determine category attribution may constitute a problem with comprehension for medical staff. Nonetheless, other classification techniques - including neural networks - and preprocessing approaches should be investigated in depth.

Another aspect is a computing cost when handling large volume of data with multivariate features which brings the necessity of good feature selection or principal component analysis [34]–[36]. Also, multi-class imbalanced problems

with at least two minority classes where the experts do not agree to aggregate them together require more advanced approaches for example with unequal costs of misclassification between classes [37].

REFERENCES

[1] Stefanowski J.: "Dealing with Data Difficulty Factors while Learning from Imbalanced Data", Challenges in Computational Statistics and Data Mining, 2016, pp. 333–363, DOI: 10.1007/978-3-319-18781-5_17.

[2] Senthilkumar D., Paulraj S.: "Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines", International Journal of Engineering and Technology, vol.5(5), 2013, pp. 3922-3929.

[3] Arslan A.K., Colaka C.: "Different medical data mining approaches based prediction of ischemic stroke", Computer Methods and Programs in Biomedicine, 2016, vol. 130, pp. 87–92, DOI: 10.1016/j.cmpb.2016.03.022.

[4] Wosiak A., Dziomdziora A.: "Feature Selection and Classification Pairwise Combinations for High-dimensional Tumour Biomedical Datasets", Schedae Informaticae, 2015, vol. 24, pp. 53-62, DOI: 10.4467/20838476SI.15.005.3027.

[5] Glinka K., Wosiak A., Zakrzewska D.: "Improving Children Diagnostics by Efficient Multi-label Classification Method", Information Technologies in Medicine 2016 vol. 1, series: Advances in Intelligent Systems and Computing 471(1), eds.: Ewa Pietka, Pawel Badura, Jacek Kawa, Wojciech Wieclawek, Springer International Publishing, pp. 253-266, DOI: 10.1007/978-3-319-39796-2.

[6] Levashenko V., Zaitseva E.: "Fuzzy Decision Trees in medical decision Making Support System" 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), Wroclaw, 2012, pp. 213-219.

[7] He H., Garcia E. A.: "Learning from Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, 2009, vol. 21(8), pp. 1263–1284, DOI: 10.1109/TKDE.2008.239.

[8] Yang Q., Wu X.: "Challenging problems in data mining research", International Journal of Information Technology and Decision Making, 2006, vol. 5(4), 597–604, DOI: 10.1142/S0219622006002258.

[9] Sun Y., Wong A.K., Kamel M.S.: "Classification of imbalanced data: A review", International Journal of Pattern Recognition and Artificial Intelligence, 2009, vol. 23(4), pp. 687–719, DOI: 10.1142/S0218001409007326.

[10] Weiss G.M., Provost F.: "Learning when training data are costly: The effect of class distribution on tree induction", Journal of Artificial Intelligence Research, 2003, vol. 19, pp. 315–354, DOI: 10.1613/jair.1199.

[11] Japkowicz N.: "Learning from Imbalanced Data Sets: A Comparison of Various Strategies", In: Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets, Austin, TX, USA, 2000.

[12] de Morais R. F., Miranda P. B., Silva, R. M.: "A Meta-Learning Method to Select Under-Sampling Algorithms for Imbalanced Data Sets", In: Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on, pp. 385–390, DOI: 10.1109/BRACIS.2016.076.

[13] Morent D., Stathatos K., Lin W. C., Berthold M. R.: "Comprehensive PMML preprocessing in KNIME", In: Proceedings of the 2011 workshop on Predictive markup language modeling, 2011, pp. 28-31, DOI: 10.1145/2023598.2023602.

[14] Wilk S., Stefanowski J., Wojciechowski S., Farion K.J., Michalowski W.: "Application of Preprocessing Methods to Imbalanced Clinical Data: An Experimental Study", In: Pietka E., Badura P., Kawa J., Wieclawek W. (eds.) Information Technologies in Medicine. Advances in Intelligent Systems and Computing, 2016, vol. 471, pp. 503–516, DOI: 10.1007/978-3-319-39796-2_41.

[15] Wong, T.T.: "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", Pattern Recognition, 2015, vol. 48(9), pp. 2839-2846, DOI: 10.1016/j.patcog.2015.03.009.

[16] Yadav S., Shukla S.: "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification", In: Advanced Computing (IACC), 2016 IEEE 6th International Conference on, pp. 78-83, DOI: 10.1109/IACC.2016.25.

[17] Zhang Y., Yang Y.: "Cross-validation for selecting a model selection procedure", Journal of Econometrics, 2015, vol. 187(1), pp. 95-112, DOI: 10.1016/j.jeconom.2015.02.006.

[18] Berthold M.R., Cebron N., Dill F., Gabriel T.R., KÃűtter T., Meinl T., Ohl P., Sieb Ch., Thiel K., Wiswedel B.: "KNIME: The Konstanz Information Miner" In: Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, Heidelberg, 2008, DOI: 10.1007/978-3-540-78246-9_38.

[19] O'Hagan S., Kell D.B.: "Software review: the KNIME workflow environment and its applications in genetic programming and machine learning", Genetic Programming and Evolvable Machines, 2015, vol. 16(3), pp. 387-391, DOI: 10.1007/s10710-015-9247-3.

[20] Lopez, V., Fernandez, A., Moreno-Torres, J. G., Herrera, F.: "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics", Expert Systems with Applications, 2012, vol. 39(7), pp. 6585–6608 DOI: 10.1016/j.eswa.2011.12.043.

[21] Drummond C., Holte R.C.: "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling", In: Workshop on Learning from Imbalanced Data Sets II, International Conference on Machine Learning, Washington, DC, USA, 2003.

[22] Garcia V., Sanchez J.S., Mollineda R.A.: "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance", Knowledge-Based Systems, 2012, vol. 25(1), pp. 13–21, DOI: 10.1016/j.knosys.2011.06.013.

[23] Fernandez A., Lopez V., Galar M, Jose del Jesus M., Herrera F.: "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches", Knowledge-Based Systems, 2013, vol. 42, pp. 97–110, DOI: 10.1016/j.knosys.2013.01.018.

[24] Chawla N., Bowye K., Hall L., Kegelmeyer W.P.: "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 2002, vol. 16, pp. 321âĂŞ357, DOI: 10.1613/jair.953.

[25] Weiss G.M.: "Mining with rarity: a unifying framework", ACM SIGKDD Explorations Newsletter, 2004, vol. 6(1), pp. 7–19, DOI: 10.1145/1007730.1007734.

[26] Batista G.E., Prati R.C., Monard M.C.: "Balancing strategies and class overlapping". In: Advances in Intelligent Data Analysis VI, 2005, pp. 24–35, DOI: 10.1007/11552253_3.

[27] Ali A., Shamsuddin S.M., Ralescu A.L.: "Classification with class imbalance problem: A Review", International Journal of Advances in Soft Computing and its Applications, 2015, vol. 7(3), pp. 176–204.

[28] https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data

[29] Kent Ridge Biomedical Dataset Repository: http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Michigan.html

[30] https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/hypothyroid.data

[31] http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/ann-train.data

[32] http://www.cbioportal.org/study?id=lusc_tcga

[33] Yang P., Xu L., Zhou B. B., Zhang Z., Zomaya A. Y.: A particle swarm based hybrid system for imbalanced medical data sampling. BMC genomics, 2009, vol. 10(3):S34, DOI: 10.1186/1471-2164-10-S3-S34.

[34] Janousova E., Schwarz D., Kasparek T.: "Data reduction in classification of 3-D brain images in the schizophrenia research", Analysis of Biomedical Signals and Images, 2010, vol. 20, pp. 69–74.

[35] Pancerz K., Paja W., Gomula J.: "Random Forest Feature Selection for Data Coming from Evaluation Sheets of Subjects with ASDs", Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, 2016, Vol. 8, pages 299–302, DOI: 10.15439/2016F274.

[36] Paja W.: "Medical diagnosis support and accuracy improvement by application of total scoring from feature selection approach", Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FEDCSIS 2015), Annals of Computer Science and Information Systems, eds. M. Ganzha and L. Maciaszek and M. Paprzycki, IEEE, 2015, pp. 281–286, DOI: 10.15439/2015F361.

[37] El-Ghamrawy S. M.: "A Knowledge Management Framework for imbalanced data using Frequent Pattern Mining based on Bloom Filter", In: Computer Engineering & Systems (ICCES), 2016 11th International Conference on, pp. 226–231, DOI: 10.1109/ICCES.2016.7822004.