DATA PREPROCESSING IN THE CLASSIFICATION OF THE IMBALANCED DATA

Katarzyna Borowska[1], Magdalena Topczewska[2]

[1] Student of Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

[2] Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: The article concerns the problem of imbalanced data classification. Two algorithms improving the standard SMOTE method have been created and tested. To measure the distance between objects the Euclidean or the HVDM metric was applied, depending on the number of nominal attributes in a dataset.

Keywords: class imbalance, oversampling, classification

1. Introduction

Dynamic development of expert systems brings uncountable benefits in many domains. Especially the medical diagnosis requires very accurate and infallible computer decision support systems. Introducing these kind of systems to real–world problems revealed that effectiveness of data mining depends on data distribution. Experts have found that standard classifiers become not sufficient when processing complex examples. One of the reasons of high complexity is an imbalanced class distribution. This problem occurs when one class is underrepresented in a dataset. It may lead to the reduction of performance when standard classifiers are used. Due to the fact that many real–life domains suffer from the class imbalance problem, it has emerged as one of the challenges in data mining community [9].

There is one fundamental difficulty when objects are classified: the assumption that concerns the class distribution. Typically, the distribution of examples in a dataset is predicted to be uniform and costs of misclassification are expected to be equal for all classes. This is excessively simplified assumption, because many datasets contain rare objects which represent the class of interest. Simple classifiers tend to generalize

Advances in Computer Science Research, vol. 11, pp. 31-46, 2014.

and create rules with broader data coverage. Hence, they usually ignore rare examples [12]. However, the accuracy may reach very high values, even when all instances from the minority class are misclassified. Therefore more appropriate methods for evaluating classification performance in case of occurrence of rare examples need to be applied.

It should be emphasized that the erroneous recognition of examples from the minority class may lead to disastrous consequences. Medical diagnosis, detection of fraudulent financial transactions, anomaly detection, learning word pronunciation, predicting pre-term births or detection of oil spills are only a few examples of areas affected by the imbalanced data problem [1,9]. It is obvious that the necessity of minimizing the number of wrong decisions in these domains is recognized as the significant issue. Due to the critical role of imbalanced data in the modern world, many proposals have been developed to decrease the negative effects of this problem. These techniques can be divided into three main categories: data level, algorithm level and cost-sensitive approaches.

Data level approaches are the most versatile. These algorithms are used independently of the classifier and that is considered as their main advantage. The principal aim of designing these solutions was to reduce disparity between the number of data from the minority and majority class. The ability to deal with additional difficulties of the learning process is very important in data processing.

One of the proposals concerning pre–processing imbalanced data is SMOTE. Although this technique has been used with success in many domains, it is not deprived of some drawbacks. Limitations of this algorithm may lead to considerable depletion of classifier performance. This problem usually occurs when dataset is not only imbalanced, but also have a complex distribution [4]. Complex distribution may be associated with factors such as overlapping, small disjunctions and noise. Recent studies have shown that these difficulties are the main source of problems in the classification [11,9,5].

The novel approach for mining imbalanced data is presented in this paper. Two algorithms are proposed as the improved version of the standard SMOTE technique.

2. Algorithm

Three main approaches can deal with the class imbalance problem. Although all of these solutions have been successfully applied in many domains, in this paper we focus only on the data level techniques as they are independent of the classifier and therefore flexible. Studies have shown that the application of preprocessing phase to balance the skewed class distribution usually improves the classifier performance [9].

Data preprocessing methods consist of various approaches. It is possible to list the following groups involving data level classification techniques:

- undersampling - a subset of the original dataset is created, some examples from the majority class have to be removed,
- oversampling - new examples are generated, especially from the minority class,
- hybrid - combination of the two previous methods.

Neither undersampling nor oversampling is deprived of disadvantages. The major drawback of undersampling is the risk of missing potentially important data. On the other hand, the oversampling in its simplest version, assuming random replication of minority class examples, may lead to overfitting [4]. The numerous proposals addressing these problems were developed.

SMOTE (Synthetic Minority Oversampling Technique) [6] represents the group of oversampling techniques. Unlike the simple random oversampling, SMOTE does not create new instances by generating identical copies of existing minority samples. The main idea of this algorithm is to create new minority class examples along the line segments between each positive class object and any of the k nearest neighbors. New instances are generated by randomly selecting appropriate number of the k nearest neighbors of a positive class example and creating a combination of features describing each of them and sample under consideration. The number of neighbors involved in oversampling depends on the number of needed minority examples. Algorithm 1 presents the pseudocode for SMOTE.

The objective of data preprocessing in SMOTE is to create synthetic examples regarding similarity between minority class instances. The similarity is defined in feature space by using the kNN algorithm, where the number of nearest neighbors k is a parameter. It is crucial to choose an appropriate value for the k parameter. Necessity of finding an adequate number of nearest neighbors is one of the SMOTE drawbacks. Over generalization and variance are considered as the other limitations of this method [4]. Although SMOTE avoids the overfitting problem and makes the decision boundaries for the minority class larger, the algorithm does not take into consideration the neighborhood of the minority class examples. It may lead to the overlapping between classes, which has a considerable negative impact on the classification process.

3. Methods

The method for creating synthetic items based on the combination of adjacent objects features seems to be so effective and groundbreaking tool that would be unreasonable

Algorithm 1 SMOTE (T, N, k)

```
Require: Number of minority class samples $T$;
    Amount of examples to create $N\%$;
    Number of nearest neighbors $k$
if $N < 100$ then
    Randomize the $T$ minority class samples
    $T = (N/100) * T$
    $N = 100$
end if
$N = (int)(N/100)$
$numattrs$ = number of attributes
$Sample[][]$: array for the original minority class samples
$newindex$: keeps a count of number of synthetic samples generated, initialized to 0
$Synthetic[][]$: array for synthetic samples
for $i \leftarrow 0$ to $T$ do
    Compute $k$ nearest neighbors for $i$, save the indices in the $nnarray$
    $Populate(N, i, nnarray)$
end for
/* Function to generate the synthetic samples */
$Populate(N, i, nnarray)$
while $N \neq 0$ do
    Choose a random number between 1 and $k$, call it $n$. This step chooses one of the $k$ nearest neighbors of $i$.
    for $attr \leftarrow 1$ to $numattrs$ do
        $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
        $gap = randomnumberbetween0and1$
        $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
    end for
    $newindex++$
    $N = N - 1$
end while
return
```

not to take advantage of the potential it brings. The original version of SMOTE technique turns out to be insufficient when faced with complex problems. On the other hand, it has been proven that many real data sets has a complicated structure and differences between the representatives of the different classes are not as obvious as it is expected [5,11]. When data is imbalanced, the high complexity has a negative impact mainly on identifying minority class instances. Especially small disjunctions, noise and overlapping handicap the process of classification (figure 1). Improved SMOTE (IS) algorithms were created to reduce negative impact of these impediments. The novel algorithms comprise a compound of the existing approaches and provide a brand new way of dealing with imbalanced data issue. Works such as [11,7,8] pose an effective solutions for the imbalanced data problem. However, they were not sufficient for all kinds of specific domains. Therefore IS techniques were designed to be more flexible.

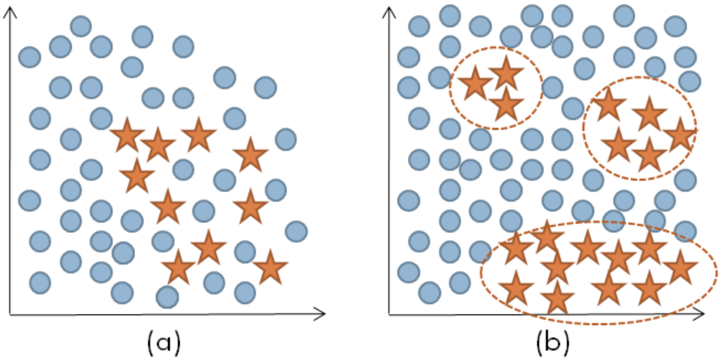


Fig. 1. (a) Class overlapping (b) Small disjunctions

Both of the developed algorithms are based on the same main concept (algorithm 27). At the beginning, right after loading the data, the metrics is choosen automatically. The analysis of attributes characteristics indicates whether the Euclidean distance is used or the HVDM. The HVDM metrics is applied, when more than half of the attributes is nominal. Otherwise, the Euclidean distance is used. The idea presented in [10] was the inspiration for developing this method to determine which distance function is the most proper for a specific issue. According to this work, the effectiveness of HVDM metrics should be closely related to the number of nominal attributes.

In the next step the k–NN algorithm is used to obtain the distance between each minority example and all other instances from both classes. The *k* is a parameter – user can specify the value of nearest neighbors. According to these calculations minority objects are divided into three groups (Algorithm 2):

- NOISE, when all of the *k* nearest neighbors represent the majority class,
- DANGER, if half or more than half of the *k* nearest neighbors come from the majority class,
- SAFE, when more than half of the *k* nearest neighbors represent the same class as the example under consideration.

Algorithm 2 IBA (S, M, k)

```
Require: Number of all instances $S$;
    Number of minority class samples $M$;
    Number of nearest neighbors $k$
metrics: keeps the name of used evaluation metric
numattrs: keeps the number of attributes
SampleMinority[][]: array for the original minority class samples
Sample[][]: array for all samples
Synthetic[][]: array for the new examples
nominal: keeps the number of nominal attributes
continuous: keeps the number of linear attributes
label[]: array for examples labels
for $i \leftarrow 0$ to numattrs do
    Verify the number of two kinds of attributes: nominal and linear. Save the number of nominal
    attributes in nominal variable and the number of linear attributes in continuous variable.
end for
if $continuous \leqslant numeric$ then
    $metrics := HVDM$
else
    $metrics := Euklides$
end if
for $i \leftarrow 0$ to $M$ do
    Calculate the distance between minority class examples and all other examples using k–NN
    method with measure written in metrics variable.
    Indexes of $k$ nearest neighbors write in nnarray array.
    $label[i] := LabelMinorityData(nnarray, i, k)$
end for
Calculate the needed number of minority class examples to create. The result save in $N$ variable.
for $i \leftarrow 0$ to $M$ do
    if $label \neq NOISE$ then
        Run the k–NN algorithm for the object $i$ using distance measure saved in metrics variable,
        indexes save in nnarray array
        $Populate(N, i, nnarray, label[i])$
        /* The Populate method is different for the two algorithms */
    end if
end for
```

The mechanism of the above division considers the location of each minority example in the feature space. There are plenty of proposals in the literature regarding this approach. They vary in the way of distinguishing different objects types and processing them in the next phases. The method presented in this paper assumes that the NOISE examples are surrounded only by the majority class instances. It is obvious that this kind of rare data may lead to serious difficulties in the learning process [2]. Examples which occure in the area surrounding class boundaries are labeled DANGER. The relatively homogeneous areas consists of the SAFE objects. The main process, responsible for generating new data, strictly relates to the assigned labels.

Algorithm 3 LabelMinorityData (nnarray, i, k)

```
Require: Number of nearest neighbors $k$;
    Indexes of $k$ nearest neighbors $nnarray$;
    Index of the example under consideration $i$;
/* This method labels the minority class data */
$minorityClass$: keeps the number of minority class neighbors of the $i$ example
$majorityClass$: keeps the number of majority class neighbors of the $i$ example
if $majorityClass == k$ then
    $returnNOISE$
end if
if $majorityClass < k/2$ then
    $returnSAFE$
end if
if $majorityClass \geqslant k/2$ then
    $returnDANGER$
end if
```

The oversampling techniques necessitate the number of minority class examples which should be created. In proposed IS solution this number is selected automatically. The minority class instances are generated to even the amount of objects from both classes.

Next, the distances between samples representing only the minority class are calculated. The k–NN method is used for this purpose again. After this operation, it is possible to start the next phase – generating new synthetic samples in the number depending on the assigned labels and the algorithm version.

The main purpose of this paper is to verify the impact of DANGER examples on the learning process. These borderline instances determine the boundaries between different classes. The two approaches of preprocessing minority samples are proposed.

3.1 ASIS

ASIS (Amplify SAFE Improved SMOTE) is the first of presented algorithms. This is a modified version of the standard SMOTE technique. The main assumption in the SIS method is that excessive number of borderline minority examples may increase the data complexity. After the steps described previously, new data is generated. Dependency between assigned labels and the number of created minority class examples is the following:

- SAFE – the numerous new objects are created, similarly to the standard SMOTE algorithm,
- DANGER – only one new example is created by combining features of the nearest neighbor and the instance under consideration; it is located closer to the DANGER object,
- NOISE – no new example is created.

The fact that the NOISE instances are omitted in this phase is dictated by their location in the feature space. Figure 2 illustrates the undesirable consequences of applying the SMOTE technique to this kind of instances. Figure 2 (a) presents the situation when the new object is created in the line segment between minority class example $A1$, considered as noisy, and its nearest neighbor $A2$, located in the homogeneous area. As it is showed, generated object $A3$ overlaps with the majority class example. The learner's ability to generalize may cause the misclassification of the $A3$. Situation is even worse in the figure 2 (b). When four neighbors are involved in creating new samples, the level of distribution disturbances is very high. The boundaries between classes are ambiguous. Hence, discriminative rules are hard to prepare. The avoidance of these problems is possible by omitting NOISE examples in preprocessing step.

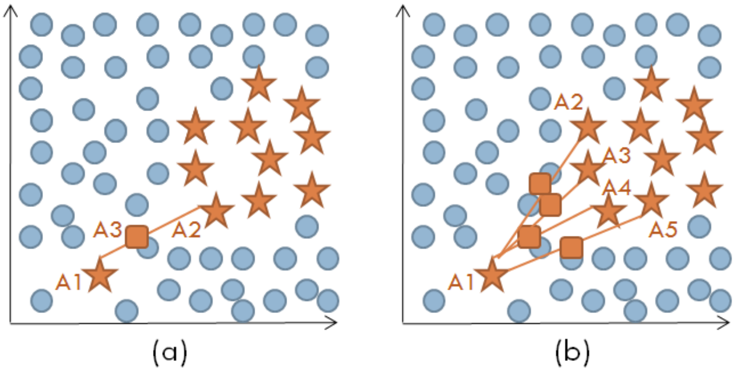


Fig. 2. Example of difficulties in processing NOISE instances

The ASIS algorithm doubles the number of DANGER data. It is assumed that this amount would be sufficient to make borderline minority examples more recognizable and not increase the overlapping level. Moreover, in this case synthetic objects are created closer to instance which is currently processed and only the nearest neighbor takes part in processing. On the other hand, the plenty of new data is created for SAFE objects. This kind of objects should be considered as the representatives of the minority class. Due to the fact that SAFE examples are located in relatively homogeneous areas this data has characteristic properties of the minority class.

3.2 ADIS

The second algorithm is named ADIS (Amplify DANGER Improved SMOTE). It represents different concept of treating respective groups of minority data. The novel strategy assumes that the ambiguity of class boundaries should be reduced by amplifying the minority class instances especially in these areas. Experiments made in [3] demonstrates that the local density of examples plays a key role in the analysis of overlapping regions containing examples from different classes. Hence, performance of the classifier should be improved when creating many new instances in DANGER objects neighborhood. In this strategy, the following processing is performed for the respective three groups:

- SAFE – one new object is created by interpolation of the example under consideration and its nearest neighbor,
- DANGER – the numerous minority instances are created, the synthetic example is placed closer to the object under consideration,,
- NOISE – no new example is created.

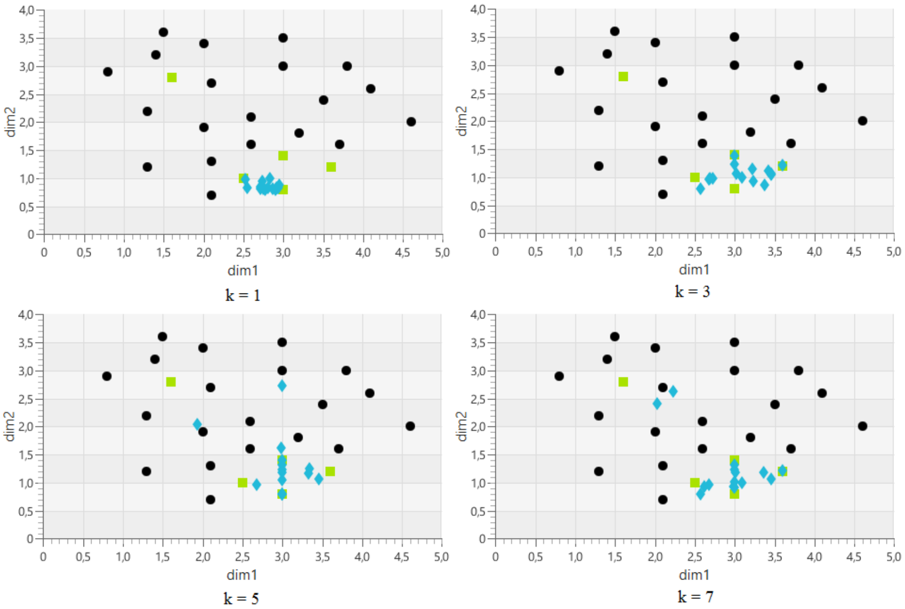
The amplification of the minority class representatives in the borderline regions should make learner to create more proper rules. However, it may lead to the degradation of the classifier performance regarding majority class examples.

4. Experiments

Two experiments have been performed to test the new methods as the improvement of the SMOTE algorithm.

4.1 Experiment 1

In the first experiment the artificial data set, containing only 25 objects, has been used to present the performance of new methods. This data are characterized by the



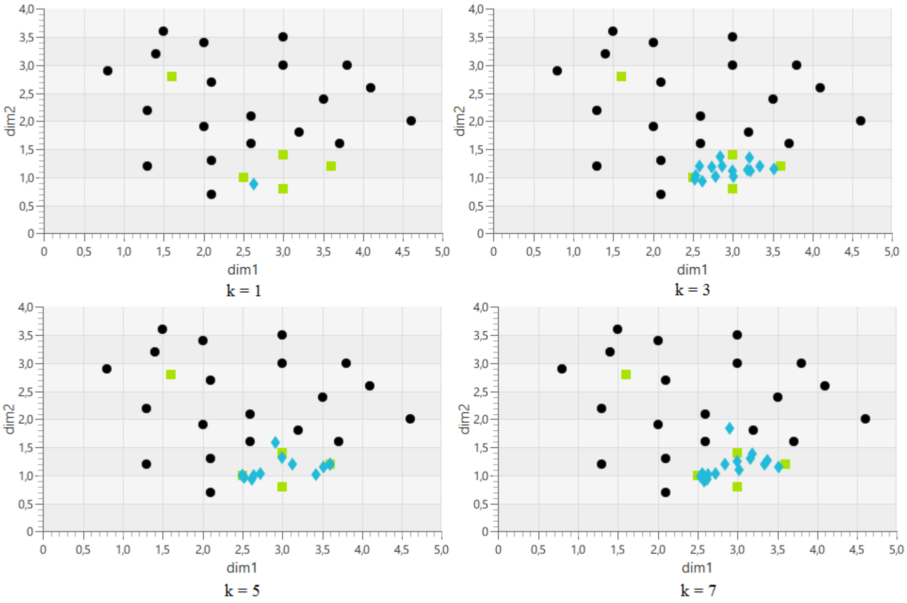
k = 1

k = 3

k = 5

k = 7

Fig. 3. Distributions after ASIS preprocessing



k = 1

k = 3

k = 5

k = 7

Fig. 4. Distributions after ADIS preprocessing

moderate IR value. The size of the majority class four times exceeds the number of objects in the minority class. The obtained results of classification are shown in the table 1. In the picture 3 and 4 the majority class objects are marked as circles, while the minority class objects are squares. The new generated objects are marked as diamonds.

Table 1. Results of the `artifficialData` classification (*IR*=4): Q – accuracy, TP_{rate} – rate of true positives, TN_{rate} – rate of true negatives, AUC – area under the curve, *nIR* – new value of IR

method	k	Q	TP_{rate}	TN_{rate}	*F-measure*	AUC	*nIR*
SMOTE	1	80.00	80.00	80.00	0.80	80.00	1.00
ASIS	1	80.00	90.00	70.00	0.82	80.00	1.00
ADIS	1	84.62	50.00	95.00	0.60	72.50	3.33
SMOTE	3	87.50	90.00	85.00	0.88	87.50	1.00
ASIS	3	95.00	90.00	100.00	0.95	95.00	1.00
ADIS	3	95.00	90.00	100.00	0.95	95.00	1.00
SMOTE	5	80.00	80.00	80.00	0.80	80.00	1.00
ASIS	5	90.00	85.00	95.00	0.89	90.00	1.00
ADIS	5	95.00	90.00	100.00	0.95	95.00	1.00
SMOTE	7	80.00	80.00	80.00	0.80	80.00	1.00
ASIS	7	75.00	65.00	85.00	0.72	75.00	1.00
ADIS	7	87.50	85.00	90.00	0.87	87.50	1.00

The analyzed data set may be specified as implicitly imbalanced, because the size of the minority class is extremely small. The best results have been obtained for the number of neighbours equaled 3. The ASIS and the ADIS methods occurred to give the highest values of parameters: the accuracy at the level 95%, the rate of true positives (90%), the rate of true negatives (100%), the F-measure (0.95) and the area under the curve (the level 95%). The same highest results have been gained for the number of neighbours 5. Thus we confirmed the assumption that the correctness of the classification depends largely on the complexity of the data distribution. Placement of objects from the minority class in a homogeneous area is one of the main success factors in creating the correct model.

4.2 Experiment 2

In the second experiment the data sets from the UCI (*University of California at Irvine Repository*) [13] are performed. The characteristics of the chosen data is presented in the table 2.

Table 2. Characteristics of datasets

dataset	number of objects	number of attributes (numeric; symbolic)	missing data	*IR*
abalone9-18	731	8 (7;1)	no	16.60
blood transfusion	748	4 (4;0)	no	3.20
breast cancer	286	9 (0;9)	yes	2.36
german credit	1000	20 (7;13)	no	2.33
hepatitis	155	19 (6;13)	yes	3.84
vowel0	988	13 (13;0)	no	9.98

Six chosen datasets have been preprocessed using the ASIS and ADIS methods. The detailed results of classification are demonstrated in the tables 3 and 4 regarding several values for the number of neighbours. The highest values are highlighted in bold.

In the case of proposed methods, both approaches proved to be effective in the real applications. For all chosen datasets the accuracy, the area under the curve, the F-measure and what is the most important – the rate of true positives are better than in the standard SMOTE case.The 3 neighbours occurred the most advantageous.

5. Conclusions

In the era of collecting increasingly large and large data volumes, the problem of the class imbalance in data becomes one of the biggest challenges for the scientists. Achieving high classification accuracy of data representing the minority class is not an easy task and the variety of methods created only for this purpose may be the confirmation of that growing need.

Among many, the algorithms belonging to the group called pre-processing of data can be noticed. Their aim is to increase in the number of objects of the positive (minority) class. The most well-known technique of this type is the SMOTE algorithm that became the inspiration to create and test the new, improved versions of the method.

In the first algorithm (`ASIS`) the number of safe objects increases primarily. They may be perceived as the best representation data of the minority class. Additionally, the number of border objects is doubled, while the objects recognized as the noise do not transform. In the second algorithm (`ADIS`) most of the new objects are created in the border area. The number of the safe objects is doubled, while similarly as in the previous case the noise objects do not share the transformation process. Comparing

Table 3. Results of the chosen UCI datasets classification: Q – accuracy, TP_{rate} – rate of true positives, TN_{rate} – rate of true negatives, AUC – area under the curve, nIR – new value of IR

method	k	Q	TP_{rate}	TN_{rate}	*F-measure*	AUC
`abalone9-18`						
SMOTE	-	94.12	35.71	97.68	0.41	0.6670
ASIS	3	97.31	**96.81**	97.82	**0.97**	97.31
	5	96.73	96.37	97.10	**0.97**	96.73
	7	**97.39**	96.52	**98.26**	**0.97**	**97.39**
	11	93.80	62.32	96.95	0.65	79.64
	15	94.48	62.50	97.82	0.68	80.16
ADIS	3	97.02	**97.10**	96.95	0.97	97.02
	5	96.73	96.37	97.10	0.97	96.73
	7	96.08	96.37	95.79	0.96	96.08
	11	95.07	95.50	94.63	0.95	95.07
	15	94.27	95.21	93.32	0.94	94.27
`blood transfusion`						
SMOTE	-	76.07	33.71	**89.30**	0.40	61.50
ASIS	3	**80.88**	79.65	82.11	**0.81**	**80.88**
	5	79.65	77.37	81.93	0.79	79.65
	7	79.91	77.72	82.11	0.79	79.91
	11	76.05	77.72	74.39	0.76	76.05
	15	78.85	78.07	79.82	0.79	78.95
ADIS	3	80.61	**84.04**	77.19	**0.81**	80.61
	5	77.98	80.53	75.44	0.79	77.98
	7	74.47	80.53	68.42	0.76	74.47
	11	76.49	73.86	79.12	0.76	76.49
	15	78.07	74.56	81.58	0.77	78.07
`breast cancer`						
SMOTE	-	69.50	41.18	**81.59**	0.45	61.38
ASIS	3	**69.90**	72.14	67.66	0.71	69.90
	5	70.15	70.15	70.15	0.70	70.15
	7	73.38	74.13	72.64	0.74	73.38
	11	68.66	72.14	65.17	0.70	68.66
	15	69.15	71.14	67.16	0.70	69.15
ADIS	3	74.63	**76.62**	72.64	**0.75**	**74.63**
	5	73.63	75.62	71.64	0.74	73.63
	7	69.40	70.15	68.66	0.70	69.40
	11	72.78	68.82	76.12	0.70	72.47
	15	69.65	70.65	68.66	0.70	69.65

Table 4. Results of the chosen UCI datasets classification: Q – accuracy, TP_{rate} – rate of true positives, TN_{rate} – rate of true negatives, AUC – area under the curve, nIR – new value of IR

method	k	Q	TP_{rate}	TN_{rate}	*F-measure*	AUC
			`german credit`			
SMOTE	-	69.60	47.33	79.14	0.48	63.24
ASIS	3	**80.64**	**81.29**	**80.00**	**0.81**	**80.64**
	5	79.50	80.14	78.86	0.80	79.50
	7	78.57	80.00	77.14	0.79	78.57
	11	79.00	80.29	77.71	0.79	79.00
	15	79.14	80.29	78.00	0.79	79.14
ADIS	3	80.14	80.86	79.43	0.80	80.14
	5	78.21	78.14	78.29	0.78	78.21
	7	78.64	79.00	78.29	0.79	78.64
	11	79.00	79.86	78.14	0.79	79.00
	15	79.00	79.86	78.14	0.79	79.00
			`hepatitis`			
SMOTE	-	85.81	53.13	**94.31**	0.61	73.72
ASIS	3	89.84	93.50	86.18	0.90	89.84
	5	89.02	92.68	85.37	0.89	89.02
	7	87.80	89.43	86.18	0.88	87.80
	11	91.87	**96.75**	86.99	**0.92**	91.87
	15	88.62	91.06	86.18	0.89	88.62
ADIS	3	91.46	92.68	90.24	**0.92**	91.46
	5	**92.28**	93.50	**91.06**	**0.92**	**92.28**
	7	91.87	94.31	89.43	**0.92**	91.87
	11	88.21	91.87	84.55	0.89	88.21
	15	86.99	89.43	84.55	0.87	86.99
			`vowel0`			
SMOTE	-	99.49	96.67	99.78	0.97	98.22
ASIS	3	99.78	**100.00**	99.55	**1.00**	99.78
	5	**99.94**	**100.00**	**99.89**	**1.00**	**99.94**
	7	99.83	99.89	99.78	**1.00**	99.83
	11	99.83	99.89	99.78	**1.00**	99.83
	15	99.78	99.89	99.67	**1.00**	99.78
ADIS	3	99.54	98.33	99.78	0.99	99.06
	5	99.61	99.67	99.55	**1.00**	99.61
	7	99.67	99.67	99.67	**1.00**	99.67
	11	99.67	99.89	99.44	**1.00**	99.67
	15	99.72	99.78	99.67	**1.00**	99.72

to the standard SMOTE algorithm, new methods improved the classification results like the accuracy, the rate of true positives or the area under the curve.

References

[1] G. M. Weiss, Mining with Rarity: A Unifying Framework, SIGKDD Explor. Newsl., Springer Berlin Heidelberg, 6(1), 7–19, 2004.

[2] S. Barua, Md. M. Islam, K. Murase, A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning, Neural Information Processing, Springer Berlin Heidelberg, 7063, 735–744, 2011.

[3] V. Garcia, R. A. Mollineda, J. S. Sanchez, On the k–NN performance in a challenging scenario of imbalance and overlapping, Pattern Analysis and Applications, Springer-Verlag, 11, 269–280, 2008.

[4] H. He, E. A. Garcia, Learning from Imbalanced Data, IEEE Trans. on Knowl. and Data Eng. on 21(9), 1263–1284, 2009.

[5] J. Taeho, N. Japkowicz, Class Imbalances Versus Small Disjuncts, SIGKDD Explor. Newsl. on 6(1), 40–49, 2004.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Int. Res. on 16(1), 321–357, 2002.

[7] S. Hu, Y. Liang, L. Ma, Y. He, MSMOTE: Improving Classification Performance When Training Data is Imbalanced, Computer Science and Engineering, 2, 13–17, 2009.

[8] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Int. Res. on 16(1), 321–357, 2002.

[9] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42(4), 463–484, 2012.

[10] G. E. A. P. A. Batista and D. F. Silva, How k-Nearest Neighbor Parameters Affect its Performance, Argentine Symposium on Artificial Intelligence, 1–12, 2009.

[11] K. Napierała, J. Stefanowski, S. Wilk, Learning from Imbalanced Data in Presence of Noisy and Borderline Examples, Proceedings of the 7th International Conference on Rough Sets and Current Trends in Computing, Springer-Verlag, Warsaw, 2010.

[12] Y. Sun, M. S. Kamela, A. K. C. Wongb, Y. Wangc, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition, 40(12), 3358—3378, 2007.

[13] UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/, (20.05.2014).

PRZETWARZANIE WSTĘPNE W PROBLEMIE KLASYFIKACJI DANYCH NIEZRÓWNOWAŻONYCH

Streszczenie: Artykuł dotyczy problemu klasyfikacji w przypadku, gdy mamy do czynienia z klasami niezrównoważonymi. W tym celu stworzone zostały dwa algorytmy poprawiające wyniki uzyskiwane za pomocą standardowego algorytmu SMOTE. Do pomiaru odległości między obiektami zastosowano metrykę euklidesową lub metrykę HVDM, w zależności od liczby cech nominalnych w zbiorze.

Słowa kluczowe: klasy niezrównoważone, tworzenie nowych obiektów, klasyfikacja

Artykuł w części zrealizowano w ramach pracy badawczej S/WI/2/2013.