Rough Sets in Imbalanced Data Problem: Improving Re–sampling Process

Katarzyna Borowska, Jaroslaw Stepaniuk

Rough Sets in Imbalanced Data Problem: Improving Re–sampling Process

Katarzyna Borowska, Jarosław Stepaniuk

Faculty of Computer Science
Bialystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
{k.borowska,j.stepaniuk}@pb.edu.pl
http://www.wi.pb.edu.pl

Abstract. Imbalanced data problem is still one of the most interesting and important research subjects. The latest experiments and detailed analysis revealed that not only the underrepresented classes are the main cause of performance loss in machine learning process, but also the inherent complex characteristics of data. The list of discovered significant difficulty factors consists of the phenomena like class overlapping, decomposition of the minority class, presence of noise and outliers. Although there are numerous solutions proposed, it is still unclear how to deal with all of these issues together and correctly evaluate the class distribution to select a proper treatment (especially considering the real–world applications where levels of uncertainty are eminently high). Since applying rough sets theory to the imbalanced data learning problem could be a promising research direction, the improved re–sampling approach combining selective preprocessing and editing techniques is introduced in this paper. The novel technique allows both qualitative and quantitative data handling.

Keywords: Data preprocessing, Class imbalance, Rough Sets, SMOTE, Oversampling, Undersampling.

1 Introduction

With the growing interest of knowledge researchers and increasing number of proposed solutions, the imbalanced data problem becomes one of the most significant and challenging issues of the last years. The main reason of this particular attention given to the underrepresented data is the fundamental importance of untypical instances. Considering medical diagnosis, it is obvious that the cost of not recognizing patient that suffers from a rare disease might lead to serious and irreversible consequences. Apart from this example, there are numerous domains in which imbalanced class distribution occurs, such as [14], [21], [25]: fraudulent credit card transactions, detecting oil spills from satellite images, network intrusion detection, financial risk analysis, text categorization and information filtering. Indeed, the wide range of problem occurrences increases its significance and explains the efforts put into finding effective solution.

Initially, the major cause of the classifier performance depletion was merely identified with not sufficient number of examples representing minority data. However, recent comprehensive studies carried on the nature of imbalanced data revealed that there are other factors contributing to this undesirable effects [7], [14], [21]. Small disjuncts [13], class overlapping [15] and presence of noise as well as outliers [25] were especially considered as the most meaningful difficulties. Despite lots of suggested solutions (discussed briefly in the next section), there are still many open issues, particularly regarding the flexibility of proposed methods and tuning their parameters. We decided to focus on the data–level approach as it is classifier–independent and therefore more universal. However, it is worth to mention that besides this kind of concept, there are also numerous algorithm–level and cost–sensitive methods [11].

Although many re–sampling methods were proposed to deal with imbalanced data problem, only few incorporate the rough set theory. The standard rough sets approach was developed by Pawlak (1926-2006) in 1982 [16]. Objects characterized by the same information (identical values of the provided attributes) are treated as indiscernible instances [17], [18], [23]. Hence, the idea of indiscernibility relation is introduced. Since the real–life problems are often vague and contains inconsistent information, the rough (not precise) set concept can be replaced by a pair of precise concepts, namely the lower and upper approximations. We claim that this methodology could be very useful both in preprocessing step and cleaning phase of the algorithm (see [3]). Especially the extended version, which allows the continuous values of attributes by involving the similarity relation. Therefore, we propose the adjusted VIS_RST algorithm, dedicated for both qualitative and quantitative data. What is more, new mechanisms for careful oversampling are introduced.

2 Related works

In this paper we focus only on the data–level methods addressing imbalanced data problem, as it was declared in the previous section. This category consists of classifier–independent solutions which transform the original data set into less complex and more balanced distribution using techniques such as oversampling and undersampling [11], [21]. The major algorithm representing this mentioned group is Synthetic Minority Oversampling Technique [6]. Since the approach of generating new minority samples based on the similarities between existing minority examples became very successful and powerful method, it was an excellent inspiration for researchers to develop numerous extensions and improvements. Some of them properly reduce the consequences of SMOTE drawbacks, such as over–generalization and variance [11]. Preparing the overview of related techniques, two main subjects were considered. Firstly, methods which handle additional difficulty factors are discussed. Secondly, we show the applications of rough set notions in imbalance data problem.

2.1 SMOTE–based methods dealing with complex imbalanced data distribution

Highly imbalanced datasets, especially characterized by the complex distribution, require dedicated methods. Even such groundbreaking algorithm as SMOTE turns out as insufficiently effective in some specific domains. Indeed, the most recent researches revealed that technique of dividing minority data into some categories that reflect their local characteristics is a proper direction of development. The main reason of this conclusion is the nature of real–world data. Assuming that minority class instances placed in relatively homogeneous regions of feature space are named safe, we should consider the fact that these safe examples are uncommon in real–life data sets [21]. In order to deal with complex imbalanced data distributions, many sophisticated methods were proposed:

- MSMOTE (Modified Synthetic Minority Oversampling Technique) [12] - the strategy of generating new samples is adapted to the local neighbourhood of the instance. Safe objects are processed similarly as in standard SMOTE technique. For border instances only the nearest neighbour is chosen. Latent noise representatives are not taken into consideration in undersampling.
- Borderline-SMOTE [9] - method that strengthens the area of class overlapping. Only borderline examples are used in processing.
- SMOTE-ENN [2] - technique combining oversampling and additional cleaning step. Standard SMOTE algorithm is enriched by the Edited Nearest Neighbour Rule (ENN) approach, which removes examples from both classes as long as they are misclassified by their three nearest neighbours.
- Selective Preprocessing of Imbalanced Data (SPIDER) [15] - method that identifies noisy minority data using the k–NN technique and continues processing in a way depending on the selected option: weak, relabel or strong. Chosen condition determines if only minority class examples are amplified or also majority objects are relabeled. After oversampling, noisy representatives of majority class are removed.
- Safe–Level SMOTE [5] - the algorithm applies k–NN technique to obtain the safe levels of minority class samples. New synthetic instances are created only in safe regions in order to improve prediction performance of classifiers.

2.2 Rough sets solutions for imbalanced data problem

The occurrence of noisy and borderline examples in real–domain data sets is the fact that need to be be acknowledged in most cases. Hence, the relevancy of methods dealing with these additional difficulties should be emphasized. The rough set notions appears as a promising approach to reduce data distribution complexity and understand hidden patterns in data. Before describing existing algorithms based on the rough sets approach, basic concepts of this theory are introduced.

Let U denote a finite non-empty set of objects, to be called the universe. The fundamental assumption of the rough set approach is that objects from a

set U described by the same information are indiscernible. This main concept is source of the notion referred as indiscernibility relation $IND \subseteq U \times U$, defined on the set U. The indiscernibility relation IND is an equivalence relation on U. As such, it induces a partition of U into indiscernibility classes. Let $[x]_{IND} = \{y \in U : (x, y) \in IND\}$ be an indiscernibility class, where $x \in U$. For any subset X of the set U it is possible to prepare the following characteristics [18]:

- the lower approximation of a set X is the set of all objects that can be certainly classified as members of X with respect to IND :

$$\{x \in U : [x]_{IND} \subseteq X\}, \tag{1}$$

- the boundary region of a set X is the set of all objects that are possibly members of X with respect to IND :

$$\{x \in U : [x]_{IND} \cap X \neq \varnothing \ \&[x]_{IND} \nsubseteq X\}. \tag{2}$$

The most known preprocessing methods directly utilizing the rough set theory are the following:

- filtering techniques (relabel and remove) [22] - depending on the method: majority class examples belonging to the boundary region (defined by the rough sets) are either relabeled or removed,
- SMOTE–RSB$_*$ [19] - method combining SMOTE algorithm with rough set theory by introducing additional cleaning phase of processing.

3 Proposed Algorithm VISROT - Versatile Improved SMOTE Based on Rough Set Theory

Comprehensive studies on imbalanced data problem and analysis of foregoing solutions revealed that there are many open issues and the need of more general approach dealing with wide range of different data characteristics is still actual [21]. Since most of the real–world data sets have complex distribution, researchers should pay particular attention to careful assortment of oversamping strategy [14]. In [20] two main types of SMOTE–based preprocessing algorithms are specified:

- *change–direction* methods - new instances are created only in specific areas of the input space (especially close to relatively large positive examples clusters)
- *filtering–based* techniques - SMOTE algorithm integrated with additional cleaning and filtering methods that aim to create more regular class boundaries.

The authors of this categorization claim that the first group may suffer from noisy and borderline instances. The necessity of additional cleaning phase was indicated. Since our VIS_RST [3] algorithm meets this requirement, but it is

not directly suitable for quantitative data, we decided to improve the existing approach and enable processing of any attributes' types.

The code generalization for both qualitative and quantitative data involved many adjustments and handling specific cases. The main modification concerns usage of weaker similarity concept instead of the strict indiscernibility relation [10]. We applied the HVDM distance metrics [26] as a generator of similarity measure.

The algorithm flexibility is obtained by two approaches dedicated to different types of problems. Analysis of local neighbourhood of each example enables to evaluate the complexity of data distribution. Based on the studies from [15] we assume that the occurrence of 30% of borderline and noisy instances indicates that the problem is difficult. Identification of these specific examples is performed by applying the k–NN algorithm. Continuing the categorization introduced in VIS algorithm[4], we distinguish between three types of objects, namely SAFE, DANGER and NOISE. SAFE examples are relatively easy to recognize, they are main representatives of minority class. DANGER instances are placed in the area surrounding class boundaries, they typically overlap with majority class examples. NOISE instances are rare, probably incorrect, individuals located in areas occupied by the majority class objects. The mechanism of categorization into mentioned groups is described below.

Let $DT = (U, A \cup \{d\})$ be a decision table, where U is a non-empty set of objects, A is a set of condition attributes and d is a decision attribute and $V_d = \{+, -\}$. The following rules enable labeling minority data $X_{d=+} = \{x \in U : d(x) = +\}$:

Definition 1. *Let $k > 0$ be a given natural number. Let $x \in X_{d=+}$ be an object from minority class. We define $Label : X_{d=+} \to \{NOISE, DANGER, SAFE\}$ as follows:*

- *$Label(x) = NOISE$ if and only if all of the k nearest neighbors of x represent the majority class $X_{d=-} = \{x \in U : d(x) = -\}$,*
- *$Label(x) = DANGER$ if and only if half or more than half of the k nearest neighbors of x belong to the majority class $X_{d=-}$ or the nearest neighbour of x is majority class representative,*
- *$Label(x) = SAFE$ if and only if more than half of the k nearest neighbors represent the same class as the example under consideration and the nearest neighbour of x is minority class representative.*

The explained approach involves three modes of processing of DT. None of them creates new samples using NOISE examples. The first one is defined below:

Definition 2. *HighComplexity mode: $DT \longmapsto DT_{balanced}$*

- *DANGER: the number of objects is doubled by creating one new example along the line segment between half of the distance from DANGER object and one of its k nearest neighbors. For nominal attributes values describing the object under consideration are replicated,*

– *SAFE: assuming that these concentrated instances provide specific and easy to learn patterns that enable proper recognition of minority samples, a plenty of new data is created by interpolation between SAFE object and one of its k nearest neighbors. Nominal attributes are determined by majority vote of k nearest neighbors' features.*

The second option is applied when most of examples belong to the relatively homogeneous areas:

Definition 3. *LowComplexity mode:* $DT \longmapsto DT_{balanced}$

– *DANGER: the most of synthetic samples are generated in these borderline areas, since numerous majority class representatives may have greater impact on the classifier learning, when there are not enough minority examples. Hence, many new examples are created closer to the object under consideration. One of the k nearest neighbor is chosen for each new sample when determining the value of numeric feature. Values of nominal attributes are obtained by the majority vote of k nearest neighbors' features,*
– *SAFE: there is no need to increase significantly the number of instances in these safe areas. Only one new object per existing minority SAFE instance is generated. Numeric attributes are handled by the interpolation with one of the k nearest neighbors. For the nominal features, new sample has the same values of attributes as the object under consideration.*

The third option is specified as follows:

Definition 4. *noSAFE mode:* $DT \longmapsto DT_{balanced}$

– *DANGER: all of the synthetic objects are created in the area surrounding class boundaries. This particular solution is selected in case especially complex data distribution, whoch do not include any SAFE samples. Missing SAFE elements indicates that most of the examples are labeled as DANGER (there are no homogeneous regions). Since only DANGER and NOISE examples are available, only generating new instances in neighborhood of DANGER objects would provide sufficient number of minority samples.*

Omitting NOISE examples in oversampling phase is explained by the idea of keeping data distribution complexity as low as possible. Generating new synthetic samples by utilisation of objects surrounded only by the majority class representatives may introduce more inconsistencies. However, there is no guarantee that objects labeled as NOISE are truly effects of errors or they are only outliers which are untypical since no other similar objects are provided in the imbalanced data set [21]. Hence, we do not remove any of these instances, but we also do not create new examples similar to them.

Even when examples considered as noise are excluded from the oversampling process, generating new samples by combining features of two chosen instances still may contribute to creation of noisy examples. Thus some filtering and cleaning mechanisms are advisable [20]. In order to resolve problem of introducing

additional inconsistencies we propose the technique of supervise preprocessing. The main idea of this approach is based on the lower approximation. After obtaining the threshold, algorithm identifies newly created objects that do not belong to the lower approximation of the minority class. The correctness of each element is obtained iteratively (by means of similarity relation rather than the strict indiscernibility relation). The expected proper number of new samples is assured by the increased limit of generated objects. The proposed solution consists of steps described in provided algorithm.

Algorithm VISROT

INPUT: All instances from both classes defined as $DT = (U, A \cup \{d\})$;
Number of minority class samples $M = card(\{x \in U : d(x) = +\}) > 0$;
Number of nearest neighbors $k \geq 3$.

OUTPUT: $DT_{balanced}$: minority and majority class instances after preprocessing.

1: **Step I:** Compare feature values of minority and majority instances. Remove all negative objects ($d = -$) identical to the positive ($d = +$) ones regarding all attributes.
2: **Step II:** Calculate the HVDM ($HeterogeneousValueDistanceMetric$) distance between each minority class example and every instance from majority class. Use $k - NN$ algorithm. Assign positive objects into categories (namely SAFE, DANGER and NOISE) considering rules specified in definition 1. Save numbers of instances belonging to each group (variables: $safeN$, $dangerN$, $noiseN$).
3: **Step III:** Select the strategy of processing utilizing the accomplished categorization as follows:
4: **if** $safeN == 0$ **then**
5: $mode := noSAFE$
6: **else if** $dangerN \geqslant 0.3 \cdot M$ **then**
7: $mode := HighComplexity$
8: **else**
9: $mode := LowComplexity$
10: **end if**
11: **Step IV:** Obtain the threshold (t) that enable to evaluate which minority instances belong to the lower approximation. The threshold is established by the analysis of the average objects distance. It is set to 0.25 of calculated average distance.
12: **Step V:** Compute the number of required minority class instances to even classes' cardinalities (N variable). The expected result should be increased by the 30%. Save the obtained number in $redundN$ variable.
13: **Step VI:** Over–sampling. Generate $redundN$ minority instances considering rules specified in Definitions 2, 3 and 4. Save the result in $syntheticSamples$ list. Randomize the order of new elements stored in list.
14: **Step VII:** From newly created minority examples ($syntheticSamples$) select N elements belonging to the lower approximation of minority class and add them to the $DT_{balanced}$. Assume that all objects that are insufficiently far apart from the majority class instances (their distance from any negative object is less than calculated threshold t) should not be included in the $DT_{balanced}$, since they belong to the rough set boundary region.

Generating redundant instances in Step VI protects from filtering out too many positive synthetic samples in cleaning phase. The method of determining additional objects number should be evaluated in the further research - the impact on the computing performance need to be especially investigated. We suggest that this number should be related to the complexity of the considered specific problem.

4 Experiments

The results of experimental study are presented in this section. We decided to compare our algorithm with five oversampling methods considered as successful in many domains. All of these techniques are described in section 2. Widely used C4.5 decision tree was chosen as a classifier, since it is one of the most effective data–mining methods [7]. Very important parameter of k–NN processing, namely k, was set to 5 as it was proven that this is the most suitable value for wide range of problems [8]. The HVDM metric was applied to measure the distances between objects, because it properly handles both quantitative and qualitative data [26].

Six data sets were selected to perform described experiments. They are highly–imbalanced real–life data sets obtained from the UCI repository [24]. All of them were firstly divided into training and test partitions to ensure that the results of fivefolds cross-validation would be correct. We used partitioned data available in the KEEL website [1]. The analyzed data sets are presented in table 1.

Table 1. Characteristics of evaluated data sets

dataset	objects	attributes	IR	boundary region
glass-0-1-6_vs_5	184	9	19.44	empty
ecoli-0-1-3-7_vs_2-6	281	7	39.14	nonempty
glass5	214	9	22.78	empty
ecoli-0-1_vs_5	240	6	11	nonempty
led7digit-0-2-4-5-6-7-8-9_vs_1	443	7	10.97	nonempty
ecoli-0-1-4-6_vs_5	280	6	13	nonempty

The existence of boundary region defined by the rough set notions is emphasized to verify the impact of data inconsistencies on the classifier performance preceded by the particular preprocessing techniques.

Table 2 presents the results of experiments. The area under the ROC curve (AUC) was used to evaluate classifier performance. This measure discloses the dependency between sensitivity (percentage of positive instances correctly classified) and percentage of negative examples misclassified.

VISROT algorithm introduced in this paper was evaluated in comparison with five other preprocessing techniques which performance was measured in [19].

Table 2. Classification results for the selected UCI datasets - comparison of proposed algorithm VISROT with five other techniques and classification without preprocessing step (noPRE).

dataset	noPRE	SMOTE	S–ENN	Border–S	SafeL–S	S–RSB$_*$	VISROT
glass016_vs_5	0.8943	0.8129	0.8743	0.8386	0.8429	0.8800	0.8943
ecoli0137_vs_2-6	0.7481	0.8136	0.8209	0.8445	0.8118	0.8445	0.8445
glass5	0.8976	0.8829	0.7756	0.8854	0.8939	0.9232	0.9951
ecoli01_vs_5	0.8159	0.7977	0.8250	0.8318	0.8568	0.7818	0.8636
led7digit02456789_vs_1	0.8788	0.8908	0.8379	0.8908	0.9023	0.9019	0.8918
ecoli0146_vs_5	0.7885	0.8981	0.8981	0.7558	0.8519	0.8231	0.8366

The results revealed that proposed method outperforms other algorithms in two cases (glass5, ecoli01_vs_5), one of whom has non–empty boundary region. For two data sets VISTROT has similar result as the most effective techniques. In the remaining two cases applying VISROT approach was slightly less beneficial than SMOTE and SMOTE–ENN or Safe–Level SMOTE and SMOTE–RSB$_*$. The experiments proved that the proposed algorithm is suitable to deal with real–life complex data distributions, even highly–imbalanced.

5 Conclusions and future research

In this paper we introduced new preprocessing method dedicated to both quantitative and qualitative attributes in imbalanced data problems. The described approach considers significant difficulties that lead to the misclassification of many minority class samples. Since not enough number of examples representing positive class is not the main reason of performance depletion, other factors were also considered. Especially occurrence of sub–regions, noise and class overlapping were examined as they indicates the high data complexity. Performed experiments confirms that oversampling preceded by the analysis of local neighborhood of positive instances is proper approach. Moreover, the need of additional cleaning step that removes the inconsistencies is emphasized. The VISROT results showed that rough set notions can be successfully applied to the imbalanced data problems.

We suggest that proposed algorithm should be adjusted to handle Big Data problems in future research. The values of minimal allowed distance defining weaken low approximation rule (threshold) can also be investigated.

Acknowledgements

This research was supported by the grant S/WI/3/2013 of the Polish Ministry of Science and Higher Education.

References

1. Alcala–Fdez J., Fernandez A., Luengo J., Derrac J., Garca S., Sanchez L., Herrera F., KEEL Data–Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple–Valued Logic and Soft Computing 17:2-3, 2011, 255–287.
2. Batista G.E.A.P.A., Prati R.C., Monard M.C., A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6(1), June 2004, 20–29.
3. Borowska K., Stepaniuk J., Imbalanced Data Classification: A Novel Re-sampling Approach Combining Versatile Improved SMOTE and Rough Sets, Lecture Notes in Computer Science 9842, Springer International Publishing, 2016, 31–42.
4. Borowska K., Topczewska M., New Data Level Approach for Imbalanced Data Classification Improvement, Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015, Springer International Publishing, 2016, 283–294.
5. Bunkhumpornpat C., Sinapiromsaran K., Lursinsap C., Safe-Level–SMOTE: Safe–Level–Synthetic Minority Over–Sampling TEchnique for Handling the Class Imbalanced Problem, Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, 2009, 475–482.
6. Chawla N.V., Bowyer K.W, Hall L.O., and Kegelmeyer W.P., SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16(1), 2002, 321–357.
7. Galar M., Fernandez A., Barrenechea E., Bustince H., Herrera F., A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 2012, 463–484.
8. Garca V., Mollineda R. A., Snchez J. S., On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal. Appl. 11(3–4), 2008, 269–280.
9. Han H., Wang W–Y, Mao B–H., Borderline–SMOTE: a new over–sampling method in imbalanced data sets learning, Proceedings of the 2005 international conference on Advances in Intelligent Computing - Volume Part I (ICIC'05), Springer-Verlag, Berlin, Heidelberg, 878–887.
10. Krawiec K., Słowiński R., Vanderpooten D., Learning Decision Rules from Similarity Based Rough Approximations, Rough Sets in Knowledge Discovery 2,Volume 19 of the series Studies in Fuzziness and Soft Computing, 1998, 37–54.
11. He H., Garcia E.A., Learning from Imbalanced Data. IEEE Trans. on Knowl. and Data Eng. 21(9), 2009, 1263–1284.
12. Hu S., Liang Y., Ma L., He Y., MSMOTE: Improving Classification Performance When Training Data is Imbalanced, Computer Science and Engineering. WCSE '09. Second International Workshop on, Qingdao, 2009, 13–17.
13. Jo T., Japkowicz N., Class imbalances versus small disjuncts. SIGKDD Explor. Newsl. 6(1), 2004, 40–49.
14. Napierała K., Stefanowski J., Types of minority class examples and their influence on learning classifiers from imbalanced data, Journal of Intelligent Information Systems, 46, 2016, 563–597.
15. Napierała K., Stefanowski J., Wilk S., 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 (RSCTC'10), Springer-Verlag, Berlin, Heidelberg, 158–167.

16. Pawlak Z., Rough Sets, International Journal of Computer and Information Sciences, 11(5), 1982, 341–356.
17. Pawlak Z., Skowron A., Rough sets: Some extensions, Information Sciences, 177(1), 2007, 28–40.
18. Pawlak Z., Skowron A., Rudiments of Rough Sets, Information Sciences, 177(1), 2007, 3–27.
19. Ramentol E., Caballero Y., Bello R., Herrera F., SMOTE-RSB$_*$: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and Information Systems, Springer-Verlag, 33(2), 2011, 245–265.
20. Saez J.A., Luengo J., Stefanowski J., Herrera F., SMOTEIPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, Information Sciences, 291, 2015, 184–203.
21. Stefanowski J., Dealing with Data Difficulty Factors While Learning from Imbalanced Data, Challenges in Computational Statistics and Data Mining, 2016, 333–363.
22. Stefanowski J., Wilk S., Rough Sets for Handling Imbalanced Data: Combining Filtering and Rule-based Classifiers. Fundam. Inf. 72(1–3), 2006, 379–391.
23. Stepaniuk J., Rough–Granular Computing in Knowledge Discovery and Data Mining, Springer, 2008.
24. UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/ (accessed 03.02.2017).
25. Weiss G.M., Mining with Rarity: A Unifying Framework, SIGKDD Explor. Newsl., 6, 2004, 7–19.
26. Wilson D.R., Martinez T.R., Improved heterogeneous distance functions, Journal of Artificial Intelligence Research, 6, 1997, 1–34.