Evaluating difficulty of multi-class imbalanced data

Mateusz Lango, Krystyna Napierala, Jerzy Stefanowski

Institute of Computing Science, Poznan University of Technology, Poznań, Poland
{mateusz.lango, krystyna.napierala, jerzy.stefanowski}@cs.put.poznan.pl

Abstract. Multi-class imbalanced classification is more difficult than its binary counterpart. Besides typical data difficulty factors, one should also consider the complexity of relations among classes. This paper introduces a new method for examining the characteristics of multi-class data. It is based on analyzing the neighbourhood of the minority class examples and on additional information about similarities between classes. The experimental study has shown that this method is able to identify the difficulty of class distribution and that the estimated minority example safe levels are related with prediction errors of standard classifiers.

Keywords: *Imbalanced data, Multiple classes, Supervised classification*

1 Introduction

Learning from class-imbalanced data has been a topic of intensive research in recent years. On one hand, several new specialized algorithms as well as data preprocessing methods have been developed; see their reviews in [3, 5]. On the other hand, a growing research interest has also been put into better understanding the imbalanced data characteristics which cause the learning difficulties [11].

Most of these works concentrate on binary imbalanced problems with a single minority class and a single majority class. This formulation is justified by focusing the interest on the most important class. If there are multiple classes, the original problem is transformed into binary one, e.g., by selecting a minority class and aggregating the remaining classes into a single one.

Nevertheless, in some situations it may be reasonable to distinguish more classes with low cardinalities. In such cases the aforementioned binarization becomes questionable. Consider for instance the medical problem of diagnosing two types of asthma (minority classes) and discerning them from healthy patients (majority class). Selecting one type of asthma as a minority class and aggregating the other one with the majority class leads to an unacceptable situation of considering ill patients as healthy. Aggregating all asthmatic patients into one minority class could be a better choice, but it still leads to the undesired loss of information about the asthma type.

Handling multiple minority classes makes the learning task more difficult as relations between classes become more complex [7, 12]. The current approaches to it are adaptations of the *one-against-all* or *one-against-one* decomposition into several binary subproblems [4]. Although the selected minority classes are preserved in these approaches, the information about internal data distributions or decision boundaries is lost, as in the original problem one class influences several neighboring classes at the same time.

Moreover, these decompositions do not consider the *mutual relations* between classes that are different for majority and minority classes. Consider for instance the aforementioned asthma learning problem. The two asthma classes are more closely related to each other, while their similarity to the majority class (healthy patients) is smaller and it should be taken into account while constructing a new approach to multiple imbalanced classes.

In our opinion, modeling the relations between classes is particularly useful for studying *data difficulty factors* in imbalanced data. Previous research on binary imbalanced data showed that local factors such as small sub-concepts, overlapping or rare case are more influential than the global imbalance ratio [6, 11]. These aspects have a significant impact on performance of learning methods [8]. Some classifiers and preprocessing methods are more sensitive to given data types than the others. Therefore, prior to designing and applying new learning methods, it is important to analyse the data characteristics. It is even more important for multi-class problems, where such approaches do not exist yet.

In [8] we have introduced a new approach to model several types of data difficulty, based on analysing the local neighbourhood of minority examples, which was successfully used to differentiate types of examples in binary problems [9, 11]. The results of that work are useful for constructing pre-processing methods or ensembles specialized for imbalanced data [11].

The main aim of this study is to introduce a new method to identify different types of minority examples in multi-class imbalanced data, which refer to data difficulty. When analysing the local neighbourhood of the given example to determine its difficulty, we take into account the class of each neighbour as in [8, 9]. However, we also exploit additional information about relations between classes of the analysed example and its neighbor. It is based on priorly defined *similarity* between these classes. To the best of our knowledge, this kind of handling similarity of classes has not been proposed yet for imbalanced data.

Summarizing our contribution, this work introduces a concept of class similarity and uses it to extend the method of identifying types of minority examples to a multi-class setting. It is then applied to analyze the difficulty of several artificial and real-world multi-class imbalanced datasets. Finally, the impact of data difficulty on learning abilities of several classifiers is experimentally studied.

2 Related Works

Multi-class imbalanced problems are not so intensively studied as its binary counterpart. There exist only few approaches; for their recent review see [10]. For instance, the new re-sampling techniques include static-SMOTE or Mahalanobis distance-based over-sampling [2]. Nearly all other approaches follow the idea of the decomposition of the multi-class imbalanced problem to a set of binary subproblems. Usually either one-against-all or one-against-one class binarization is integrated with appropriate balancing of binary samples or with specialized ensembles, see e.g. [4]. Few other algorithmic modifications are designed for specific learning algorithms, like SVM or neural networks. However, as it is pointed out in a review in [7], none of these methods takes into account both individual properties of classes and their mutual relations.

Research on data characteristics of multi-class imbalanced datasets is limited to one paper [10] only. The authors considered the categorization of minority examples into four types (safe, borderline, rare, outlier), proposed for binary problems in [9]. To adapt it to multi-class problems, they decomposed data into several binary problems using one-vs-all technique, i.e. all examples from different classes are treated equally when analysing the local neighbourhood of the minority examples to determine their difficulty. Again, no notion of mutual relations between the classes is taken into account.

3 Identification of difficulty degrees in multi-class imbalanced data

The relations between multiple imbalanced classes are more complex than in the binary versions. As discussed in [7, 12], a given class can be a minority class with respect to some classes, and at the same time the majority one to another subset of classes. When dealing with multiple classes, one may easily lose performance on one minority class while attempting to improve it at another class [10]. Moreover, the mutual relations between classes show that some minority classes can be treated as more closely related to each other than to the majority class. Current decomposition approaches, which treat all pairs of classes equally, do not reflect well these issues [10, 12]. Thus, new approaches should take into account the complexity of different relations between multiple classes.

Furthermore, data difficulty factors may appear only in some subsets of classes. For instance, the degree of class overlapping between different classes may be different. Analyzing the type of examples present in the given class distribution also strongly depends on their relations to other classes. For instance, a given example may be of a borderline type [8, 9] for certain classes and at the same time a safe example for the remaining classes. Using existing binary class approaches to estimate imbalanced data difficulty is not straightforward in case of multiple class imbalance. There is a need for a deeper insight into these complex relations and for a new and more flexible approach to analyse multi-class data difficulty factors.

3.1 Handling multiple class relations with similarity information

Modeling relations between multiple imbalanced classes can be realized by means of additional information. In this paper, following the motivations described in Section 1, we will exploit information about *similarity* between pairs of classes.

More precisely, given a certain class we need information about similarities of other classes to it. An intuition behind it is the following: if example x from a given class has some neighbors from other classes, then neighbors with higher similarity are more preferred. For instance, consider the asthma learning problem, in which two asthma classes are defined as more similar to each other than to the no-asthma class. If an example from asthma-type-1 is not surrounded only by examples from its class (which is the most preferred situation), then we would prefer it to have neighbors from asthma-type-2 class rather than from the no-asthma class. Such neighborhood would let us consider the analysed example to be safer – easier recognized as a member of its class (as it will be less prone to suffer from the algorithm bias toward the majority classes).

We assume that for each pair of classes C_i, C_j the degree of their similarity will be defined as a real valued number $\mu_{ij} \in [0; 1]$. Let us discuss its main properties. Similarity of a class to itself is defined as $\mu_{ii} = 1$. The similarity does not have to be symmetric, i.e. for some classes C_i, C_j it may happen that $\mu_{ij} \neq \mu_{ji}$.

Although the values of μ_{ij} are defined individually for each dataset, we claim that for the given minority class C_i its similarity to other minority classes should be relatively higher than to the majority classes.

The information about similarity should be provided by the user. It can either come from the domain knowledge or be acquired from a domain expert. If neither is available, we recommend for other minority classes C_g $\mu_{ig} \to 1$, while similarities to majority classes C_h should be $\mu_{ih} \approx 0$.

3.2 Data difficulty with respect to a safe level of minority examples

In our earlier research [8, 9, 11] we claimed that (1) imbalanced data difficulty factors correspond to *local data characteristics*, occurring in some sub-regions of the minority class distribution and (2) the mutual position of an example with respect to examples from other classes of both minority and majority classes influences learning classifiers. We linked these difficulty factors to different types of examples – *safe* and *unsafe* (difficult) for recognizing the minority class. Safe examples are located in the homogeneous sub-regions belonging to one class while unsafe examples are categorized into borderline, rare cases or outliers. In [8] we introduced the method of assessing the type of example by analyzing class labels of its surrounding examples. The neighborhood was constructed based either on k–nearest neighbors or on kernel functions.

In this study we consider the k–nearest neighbors variant[1]. Determining the number of examples from the majority class in the neighborhood of the minority example allows to assess how safe the example is, and then establish its type. Below we adapt this idea for the multiple imbalanced class framework.

[1] Refer to [9] for details of the neighborhood construction, recommended distance functions and neighborhood size tuning.

Considering a given example x belonging to the minority class C_i we define its safe level with respect to l classes of examples in its neighborhood as:

$$safe(x_{C_i}) = \frac{\sum_{j=1}^{l} n_{C_j} \mu_{ij}}{n}$$

where μ_{ij} is a degree of similarity, n_{C_j} is a number of examples from class C_j inside the considered neighborhood of x and n is a total number of neighbors.

Given the safe levels calculated for all learning examples, one can analyse this information in two ways: either analyze numeric distribution of safe levels in the learning set for each class, or transform the continuous safe levels into discrete intervals corresponding to types of example (as done in [8, 9]). In the next section we follow the first option and then aggregate the distributions for each class, e.g., by the average. They should be interpreted in the following way: the lower the average value, the more unsafe (difficult) is the minority class. The statistics for each minority class can be analysed independently or can be further aggregated into a single criterion describing the difficulty of the whole learning set. Alternatively, the histograms of safe levels in each class can be presented to the user.

4 Experimental evaluation

In the experiments we want to examine three aims: (1) verify whether the new approach to evaluate the safe level (see Section 3.2) sufficiently reflects the difficulty of multi-class dataset; (2) compare this approach against its binary predecessor; (3) check whether values of safe levels relate to classification performance of standard algorithms. In order to check these aims, we will use several synthetic and real-world multi-class imbalanced datasets.

The artificial datasets were constructed to control their level of difficulty [13]. They are two-dimensional with two minority classes, having elliptic shapes, surrounded by the examples from the majority class. Each data set contains 1200 examples with the class ratio 1:2:7. In the first dataset (A1), two minority classes are well separated from each other and also from the majority class (see Fig. 1). Then, it is modified to (A1b) version by introducing an overlapping border with the majority class. In the third dataset (A2) minority class ellipses are additionally overlapping (see Fig. 1). The most difficult dataset (A3) additionally contains rare cases, outliers and more borderline examples.

Following similar motivations and analysis of the previous research on difficulty of binary imbalanced data [9] we chose three UCI datasets: `new thyroid` (NT) is a safe (easy) dataset, `ecoli` (EC) is a borderline dataset, and `cleveland` (CL) is a rare/outlier dataset. Characteristics of these datasets are presented in Table 1 and their visualisations after the reduction to two dimensions using the MDS method are presented on Fig. 1. If some of these datasets contain more majority classes, we aggregated them to be consistent with the artificial data setup.

Dataset	Abbrev.	Size	Min1 name	Min1 size	Min2 name	Min2 size	Min3 name	Min3 size
new-thyroid	NT	215	2	35	3	30		
ecoli	EC	336	imU	35	om	20	pp	52
cleveland	CL	303	2	36	3	35	4	13

Table 1. Characteristics of real-world datasets

We chose three different configurations of similarity values μ_{ij} - see Table 2. In the version called `Safety1`, we set the similarity between minority classes to a high value (0.8), following the recommendation from Section 3.1. In `Safety2` we also assume high similarity between minority classes (0.7), but we assign a small similarity between majority and minority classes as 0.2. The last configuration `Safety3`, models the situation when there is no prior information about classes relation (quite small similarity between minority classes and no similarity with the majority class). The column called `Safety` refers to the previous binary class version [9].

	Safety	Safety1	Safety2	Safety3
$\mu_{min1\ min2}$	0	0.8	0.7	0.5
$\mu_{min\ maj}$	0	0	0.2	0

Table 2. Different sets of similarity values

All experiments were performed in scikit-learn or WEKA frameworks. Classification performance was evaluated in 5-fold cross validation. **Following earlier related studies, we selected CART decision tree, PART rules, Naive Bayes and 3-nearest neighbors classifier for the experiment. All clasifiers were used with default parameter values.** Values of average safe levels are presented in Table 3, while sensitivity (true-positive-rate[2]) of minority classes is given in Table 4. **It is important to note that sensitivity is reported for each class separately and no aggregation through multiple classes was used.**[3]

4.1 Analysis of artificial datasets

Let us consider the results for each dataset in the order of their increasing difficulty (from A1 to A3).

In dataset A1, classes are easily separable, so the average safe level of both minority classes is close to 1 (Table 3). It is diminished by the safe level of some

[2] **The true positive rate of a class is the number of correctly identified class examples divided by the total number of this class examples.**

[3] Artificial datasets and detailed results are available at www.cs.put.poznan.pl/mlango/publications/multi-typology.html

	Safety			Safety1			Safety2			Safety3		
	Min1	Min2	Min3	Min1	Min2	Min3	Min1	Min2	Min3	Min1	Min2	Min3
NT	0.77	0.78		0.77	0.78		0.82	0.82		0.77	0.78	
EC	0.57	0.74	0.82	0.57	0.91	0.86	0.66	0.90	0.88	0.57	0.85	0.84
CL	0.14	0.13	0.08	0.29	0.32	0.34	0.41	0.42	0.42	0.23	0.25	0.24
A1	0.91	0.96		0.91	0.96		0.93	0.97		0.91	0.96	
A1b	0.68	0.80		0.68	0.80		0.75	0.84		0.68	0.80	
A2	0.53	0.70		0.71	0.79		0.74	0.82		0.64	0.76	
A3	0.32	0.47		0.55	0.59		0.60	0.65		0.46	0.54	

Table 3. Average safe levels for real and artificial datasets.

	CART			NB			3NN			PART		
	Min1	Min2	Min3	Min1	Min2	Min3	Min1	Min2	Min3	Min1	Min2	Min3
NT	0.94	0.83		0.94	0.86		0.71	0.80		0.94	0.83	
EC	0.60	0.85	0.78	0.68	0.30	0.90	0.48	0.75	0.84	0.46	0.80	0.79
CL	0.28	0.11	0.07	0.14	0.25	0.15	0.08	0.00	0.00	0.20	0.11	0.08
A1	0.93	0.93		0.47	0.84		0.94	0.97		0.88	0.97	
A1b	0.74	0.82		0.60	0.78		0.77	0.84		0.67	0.70	
A2	0.56	0.73		0.32	0.56		0.54	0.79		0.50	0.73	
A3	0.25	0.42		0.00	0.02		0.20	0.39		0.14	0.57	

Table 4. Sensitivity of minority classes for studied classifiers.

minority examples on the border of the class, which are surrounded by majority neighbors. As in `Safety2` the degree of similarity $\mu_{min\ maj}$ is higher (0.2) than for the remaining configurations, the average safety for this configuration is also slightly increased.

A1b is similar to A1, but has higher overlapping with the majority class. Therefore, its average safe level is smaller than for A1. Similarly to A1, values for `Safety2` are higher than for other configurations (where $\mu_{min\ maj} = 0$).

Dataset A2 is a more difficult modification of A1b dataset, with additional overlapping between minority classes. It is reflected in the lower values of the average safe level compared to A1 and A1b. Let us observe, however, that in case of `Safety1`, where we defined a very high similarity between minority classes (0.8), this additional overlapping is mostly neglected, which reduces the dataset A2 to dataset A1b. It can be noticed by almost identical average safe levels of both datasets for `Safety1`.

Finally, dataset A3 is the most difficult, with additional rare cases and outliers. It has also the lowest average safe levels from all datasets, independent of the similarity degrees configuration.

For all artificial datasets, class Min1 has lower safe levels than Min2 because it is smaller, so fewer examples are placed in homogeneous safe regions.

Looking at classification performance, one can notice that the values of safe levels are related to the sensitivity of minority classes (Table 4) – dataset A1 is best recognized by all classifiers, while dataset A3 is the most difficult. Class

Min1 is always recognized worse than Min2. **Majority class recognition was always at approximately 0.9.**

To sum up, we have shown that the proposed approach is related to data difficulty – by rating the datasets from the safest (A1) to the most difficult (A3). The difficulty is also strongly related to the recognition of the minority classes by different classifiers. Moreover, it has been shown that analyzing our enhanced safeness (related to similarity degree) allows the user to differentiate overlapping of different classes, giving a better insight into the structure of the imbalanced dataset.

4.2 Analysis of real-world datasets

Looking at the MDS visualisation of NT dataset (Fig. 1), notice that all the classes are clearly separated. It is also reflected in average safe levels (Table 3). Its values of `Safety`, `Safety1`, and `Safety3` are similar, while its `Safety2` is slightly higher. This suggests that analogously to datasets A1 and A1b, the overlapping occurs only between minority and majority classes (minority classes do not overlap each other) – it is confirmed by the MDS visualisation. The results of classifiers on this dataset also show that it is of a safe type.

EC dataset is more difficult. On MDS visualisation, class Min2 partially overlaps with Min3, then Min3 overlaps with Min2 and Maj, while Min1 overlaps with Maj. It could be devised also from values of the safe levels. For Min1 the safe level is stable for `Safety`, `Safety1` and `Safety3` configurations and increases slightly for `Safety2` – which confirms that it overlaps only with majority class. The safe level of Min2, on the other hand, is the smallest for `Safety` ($\mu_{min1\ min2} = 0$) and the highest for `Safety3` ($\mu_{min1\ min2} = 0.8$) which suggests that the overlapping is mostly between minority classes. From the classification point of view, this dataset is more difficult than NT, and Min1 is the most difficult class for all classifiers (except NB). It is due to a fact that this class is surrounded by the majority class, towards which standard classifiers have a strong bias. The latter observation supports our intuition expressed in Section 3.1, that the minority neighborhood should be considered as safer (easier) than majority neighborhood, and that it should be taken into account when estimating the difficulty of the multi-class imbalanced dataset.

CL dataset is the most difficult. The MDS visualisation clearly shows that this dataset consists mostly of mixed rare and outlier examples for all classes. Its average safe level is also very low, and the classes are hardly recognized by any of the classifiers.

To sum up, the analysis on real-world datasets also shows that the proposed approach can sufficiently well estimate the difficulty of the dataset, which is consistent with both MDS visualisations and with performance of classifiers.

5 Concluding remarks

The problem of learning from imbalanced multi-class data is particularly challenging and requires more extensive research on its nature and sources of its

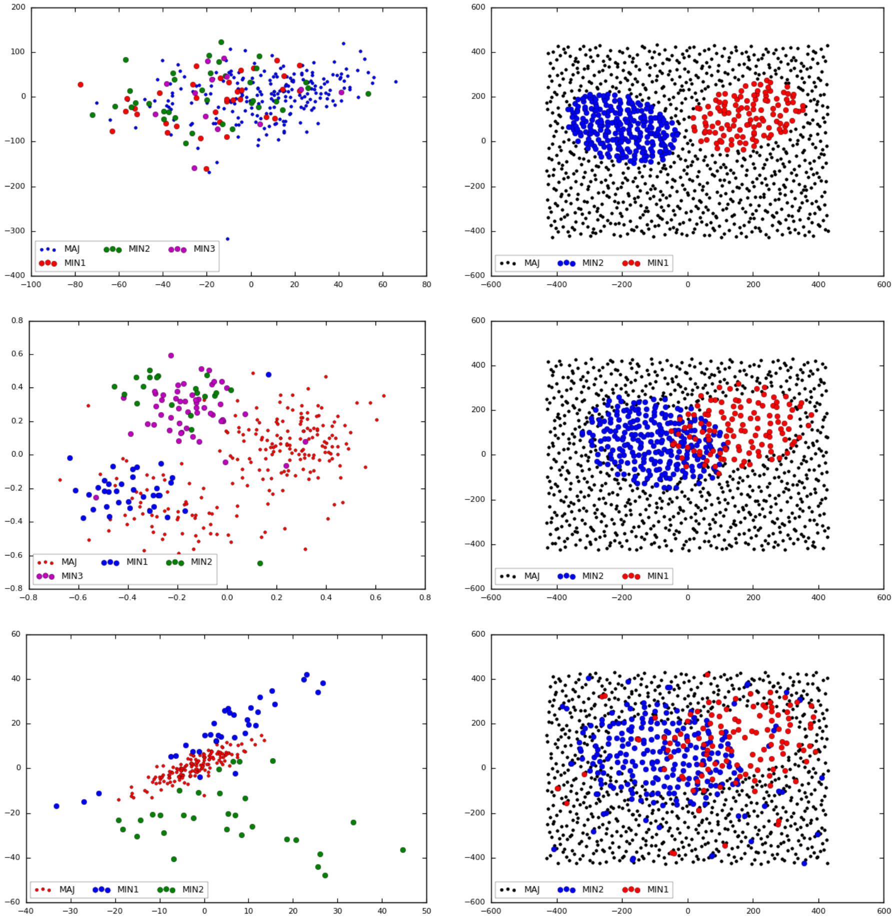

Fig. 1. MDS visualisation of studied imbalanced datasets. **In the first column from the top: CL, EC, NT; second column: A1, A2, A3.**

difficulty. In our opinion, it is necessary to analyze types of examples (safe vs. unsafe) in the distribution of the minority class. As such research is still not well-developed, we have introduced a new method. It is based both on analyzing the neighborhood of the minority class example and on the additional information about similarity of neighboring classes to the class of this example. To the best of our knowledge, similar approaches to handle complex relations among classes have not been considered yet – they were put in the main open research points of [7].

The results of experiments show that this method sufficiently identifies difficulties of minority class distributions in various artificial and real-world datasets – which is expressed by values of safe levels for appropriate minority examples. Furthermore, these values are well related to predictions of standard classifiers.

Although our method requires defining values of similarities among classes, we claim that by using them we were able to identify data difficulty factors, e.g. we could evaluate which classes overlap. Note that considering various sets of class similarities has led us to observe that the class surrounded by the majority examples is more difficult to recognize than overlapped minority classes (see an analysis of `ecoli` in Section 4.2). Experiments have also demonstrated that data difficulty factors are more influential than the global imbalance ratio.

Our proposal could also be used to construct new preprocessing methods, e.g. by exploiting safe levels to adaptively tune re-sampling. Furthermore, they could be used inside new algorithms, similarly to earlier attempts of using the local neighborhood in generalizations of under-bagging [1].

Acknowledgment. The research was funded by the the Polish National Science Center, grant no. DEC-2013/11/B/ST6/00963.

References

1. Błaszczyński, J., Stefanowski, J.: Neighbourhood sampling in bagging for imbalanced data. Neurocomputing, vol. 150 A, 184–203 (2015)
2. Abdi, L., Hashemi, S.: To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans. Knowl. Data Eng. 28(1), 238–251 (2016)
3. Branco, P., Torgo, L., Ribeiro, R.: A survey of predictive modeling under imbalanced distributions. ACM Computing Surveys (CSUR), 49(2), 31 (2016)
4. Fernandez, A., Lopez, V., Galar, M., Jesus M., Herrera, F.: Analysis the classification of imbalanced data sets with multiple classes, binarization techniques and ad-hoc approaches. Knowledge Based Systems, 42, 97–110 (2013)
5. He, H., Yungian, Ma (eds): Imbalanced Learning. Foundations, Algorithms and Applications. IEEE - Wiley, (2013)
6. Japkowicz, N., Stephen, S.: Class imbalance problem: a systematic study. Intelligent Data Analysis Journal, vol. 6 (5), 429–450 (2002)
7. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Progress Artificial Intelligence (2016) accepted for publication.
8. Napierala, K., Stefanowski, J.: The influence of minority class distribution on learning from imbalance data. In Proc. 7th Conf. HAIS 2012, LNAI vol. 7209, Springer, 139–150 (2012)

9. Napierala, K., Stefanowski, J.: Types of minority class examples and their influence on learning classifiers from imbalanced data. Journal of Intelligent Information Systems, 46(3), 563–597 (2016)
10. Seaz, J., Krawczyk, B., Wozniak, M.: Analyzing the oversampling of different classes and types in multi-class imbalanced data. Pattern Recognition, 57, 164–178 (2016)
11. Stefanowski, J.: Dealing with data difficulty factors while learning from imbalanced data. In Mielniczuk, J., Matwin, S. (eds) Challenges in Computational Statistics and Data Mining, Springer, 333–363 (2016)
12. Wang, S., Yao, X.: Mutliclass imbalance problems: analysis and potential solutions. IEEE Trans. System Man Cybern., Part B. 42 (4), 1119–1130 (2012)
13. Wojciechowski, S., Wilk, Sz.: The generator of synthetic multi-dimensional data. Poznan Univ. of Technology Report RB-16/14 (2014)