EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Over-sampling imbalanced datasets using the Covariance Matrix

Ireimis Leguen-deVarona[1], Julio Madera[1,*],Yoan Martínez-López[1] and José Carlos Hernández-Nieto[1]

[1]University of Camagüey, Camagüey, Cuba

Abstract

INTRODUCTION: Nowadays, many machine learning tasks involve learning from imbalanced datasets, leading to the miss-classification of the minority class. One of the state-of-the-art approaches to ”solve” this problem at the data level is Synthetic Minority Over-sampling Technique (SMOTE) which in turn uses K-Nearest Neighbors (KNN) algorithm to select and generate new instances.

OBJECTIVES: This paper presents SMOTE-Cov, a modified SMOTE that use Covariance Matrix instead of KNN to balance datasets, with continuous attributes and binary class.

METHODS: We implemented two variants SMOTE-CovI, which generates new values within the interval of each attribute and SMOTE-CovO, which allows some values to be outside the interval of the attributes.

RESULTS: The results show that our approach has a similar performance as the state- of-the-art approaches.

CONCLUSION: In this paper, a new algorithm is proposed to generate synthetic instances of the minority class, using the Covariance Matrix.

Received on 29 February 2020; accepted on 04 April 2020; published on 15 April 2020

Keywords: Imbalanced datasets, Oversampling, Covariance Matrix, Attribute Dependency

doi:10.4108/eai.13-7-2018.163982

1. Introduction

Covariance Matrix could be used in different continuous optimization problems[25], such as energy resource management problem(ERM)[26]. Moreover, real world data often presents characteristics that affect classification: noise, missing values, inexact or incorrect values, inadequate data size, poor representation in data sampling, etc. The imbalanced dataset problem represents a field of interest as it occurs when the number of instances that represent one class(rare events) [1] is much larger than the other classes, a common problem in certain areas such as fraud detection, cancer gene expressions, natural disasters, software defects, and risk management[2]. Rare events are difficult to detect because of their infrequency and casualness; misclassification of rare events could often results in heavy costs. For example, for smart computer security threat detection [3], dangerous connection attempts may only appear out of hundreds of thousands log records, but failing to identify a serious vulnerability breach would cause enormous losses. Moreover other examples are: the classification of the imbalanced data using radial-based undersampling [31], the learn of the imbalanced data improving interpolation-based oversampling[32], and the analysis of attribute mapping rules for recogniting in imbalaced dataset of DNA sequence applying SVM [34]. Other examples of imbalaced datasets are:

- Consistent performance of its High Voltage Circuit Breaker (HCVB) is determinate when it needs maintenance, which is an important problem, since these components are used over wide periods of time[27].

- The forecast accuracy of the ramp events tends to be low is a class imbalance problem, where take on some data sampling methods to overwhelmed[28].

[*]Corresponding author. Email: julio.madera@reduc.edu.cu

- The research of detecting anomalies in smart grid is a current topic and is investigated by many researchers, taking into account the use recognized methods of pattern recognition[29].

Some datasets for forecasting of energy is not balance[1], for example: 1) the power electric generation by electricity market module region and source; 2) the alternative fueling station locations; and 3) the electricity consumption from the California Energy Commission sorted by residential and non-residential from 2006 to 2009.

Then, in the case of the datasets with binary class, it can be defined that it is balanced if it has an approximately equal percentage of examples in the concepts to be classified, that is, if the distribution of examples by classes is uniform, otherwise it is imbalanced. To measure the degree of imbalance of a problem [4] defined the Imbalanced Ratio (IR) as:

$$IR = \frac{|C+|}{|C-|} \geq 1.5 \tag{1}$$

where:

> $C+$: Number of instances that belong to the majority class
>
> $C-$: Number of instances that belong to the minority class

Therefore, a dataset is imbalanced when it has a marked difference (IR $\geq$ 1.5) between the examples of the classes. This difference causes low predictive accuracy for the infrequent class as classifiers try to reduce the global error without taking into account the distribution of the data. In imbalanced sets, the original knowledge is usually labelled as oddities or noise, focusing exclusively on global measurements [5]. The problem with the imbalance is not only the disproportion of representatives but also the high overlap between the classes. To face this problem diverse strategies have been developed and can be divided into four groups: at the data level [6, 7], at the learning algorithms level [8], cost-sensitive learning [9] and based on multi-classifiers[10]; being the techniques at the level of the data the most used, because its use is independent of the classifier that is selected.

One of the best-known algorithms within data-level techniques is the Synthetic Minority Oversampling Technique (SMOTE) [7, 11] for the generation of synthetic instances. One of SMOTE's shortcomings is that it generalizes the minority area without regard to the majority class leading to a problem commonly know as overgeneralization; this has been solved with the use of cleaning methods such as SMOTE – Tomek links (TL) [6, 11], SMOTE - ENN [6, 11], Borderline - SMOTE1 [11, 12], SPIDER [13], SMOTE-RSB* [33], ADASYN [6] among others. These algorithms have been designed to operate with values of both discrete and continuous features for problems with imbalances in their two classes; most of them use the KNN to obtain the synthetic instances, and although this is a method that offers good results, it does not take into account the dependency relationships between attributes, which can influence on the correct classification of the examples of the minority class.

A way to obtain the dependency relation of the attributes is Probabilistic Graphical Models (PGM)[14] which represent joint probability distributions where nodes are random variables and arcs conditional dependence relationships. Generally, the PGM has four fundamental components: semantics, structure, implementation, and parameters. As part of the PGM there are Gaussian Networks that are graphic interaction models for the multivariate normal distribution [15] and some use the Covariance Matrix (CM) to analyze relationships between variables.

This paper is an extension of the proceeding Conference's article[30], where an algorithm based on SMOTE and the Covariance Matrix estimation to balance datasets with continuous attributes and binary class, exploding the dependency relationships between attributes and obtaining AUC [16] values similar to the algorithms of the state-of-the-art.

An experimental study was performed ranking two SMOTE-Cov variants, SMOTE-CovI (which generates new values within the interval of each attribute) and SMOTE-CovO (which allows some values to be outside the interval of the attributes), against SMOTE, SMOTE-ENN, SMOTE-Tomek Links, Borderline-SMOTE, ADASYN, SMOTE-RSB* and SPIDER; using 7 data-sets from the UCI repository [17] with different imbalance ratios and using C4.5 as classifier. The performance of the classifier was evaluated using AUC and hypothesis-testing techniques as proposed by [18, 19] for statistical analysis of the results.

2. Over-sampling based on the Covariance Matrix

This section introduce over-sampling based on the Covariance Matrix. First, we describe the Covariance Matrix which allow to compute variable dependency. Then, we give an overview of our proposed algorithm. Finally, we describe our experimental setup in four steps: tool, dataset selection, evaluation methodology and classifier used.

2.1. Covariance Matrix

The covariance matrix contains the covariance between the elements of a vector, where it measures the

[1] https://openei.org/datasets/

linear relationship between two variables. If the vector-column entries:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \tag{2}$$

then the covariance matrix $\sum_{ij}$ is the matrix whose (i, j) entry is the covariance

$$\sum_{ij} = \mathbf{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] \tag{3}$$

where the operator **E** denotes the expected value (mean) of its argument

$$\mu_i = \mathbf{E}(X_i) \tag{4}$$

The Covariance Matrix allows determining if there is a dependency relationship between the variables and it is also the data necessary to estimate other parameters. In addition, it is the natural generalization to higher dimensions of the concept of the variance of a scalar random variable[19].

2.2. SMOTE-Cov

The Algorithm 1 show the steps of SMOTE-Cov to balance datasets [30]. During the loading of the dataset in the first step, the algorithm expects continuous-valued attributes and a binary class. Then, it uses the formula 1 to verify whether the dataset is balanced or not. If it is imbalanced, the algorithm computes the Covariance Matrix. The Covariance Matrix allows detect the dependency relationship between attributes. Then, from the estimated covariance matrix, new synthetic instances are generated to balance the minority class. This process stops when an equilibrium between the two classes is reached. The algorithm checks that all the new values generated from the covariance are obligatorily within the interval of each attribute, in the case that some are outside the interval what is done is to take it to the minimum or maximum, making a kind of REPAIR of the value.

The computational complexity of the SMOTE-Cov in the worst case is O(n^2), which is similar to some state-of-the-art approaches, such as, SMOTE–ENN, SMOTE-RSB and ADASYN.

Algorithm 1: SMOTE-Cov steps

```
Input: Dataset $X$,inRange[Boolean]
Output: Balanced dataset $X$
Data: Dataset $X$
Step 1: Load dataset $X$;
Step 2: Compute $X$ IR using equation 1 ;
if $IR \geq 1.5$ then
    Step 3: Estimate covariance matrix using
      equation 3, this will provide us with a
      probabilistic distribution of the dataset;
    Step 4: For each attribute, a range is
      determined by it min-max value;
    while $X$ is not in equilibrium do
        Step 5: Generate new instance $y$ according
          to the covariance matrix;
        if $range \neq true$ then
            add $y$ to $X$;
        else
            for $i \leftarrow 0$ to $Y_i$ do
                if $Y_i <$ min $Y_i$ then
                    $Y_i =$ min $Y_i$ ;
                else if $Y_i >$ max $Y_i$ then
                    $Y_i =$ max $Y_i$ ;
                else
                    continue;
                end
            end
        end
    end
else
    return $X$;
end
```

3. Tools and experimental setup

The algorithm was developed using the R language because it is designed for statistical processing and has the cov() function for calculating covariance. In order to evaluate the behavior of the proposed algorithm it was compared against the state-of-the-art algorithms of oversampling data balancing; two variants are taken into account: when the attributes inside or outside of the dependence range. Seven datasets from the UCI repository were chosen with IR ≥ 1.5, see Table 1, with continuous attributes and binary class. This experiment uses 5-fold cross-validation and the data is split into two subsets: training/calibration set (80%) and test set (20%). The final result is the mean of the 5 result sets. The partitions were made using KEEL in such a way that the number of instances per class remained uniform. The partitioned datasets are available on the KEEL website [21].

Table 1. Description of the datasets used in the experiments

Dataset	Instances	Attributes	IR
ecoli2	336	7	5.4
glass-0-1-2-3_vs_4-5-6	274	9	3.20
glass1	214	9	1.81
Iris	150	4	2
newthyroid2	215	5	5.14
Pima	768	8	1.86
vehicle3	846	18	2.99

The training datasets are balanced, generating new synthetic instances from the minority class to complete the quantities of the majority class and using a sample of control test, which continues imbalanced and without any modification. The new datasets are generated from the obtained instances, using the SMOTE-Cov algorithm and a classifier is used as a mean to measure the performance using other techniques.

The classifier used for the experimental study is C4.5 (implemented in the Weka package as J48) [22], which has been referred to as a statistical classifier and one of the top 10 algorithms in Data Mining that is widely used in imbalanced problems [4].

The Area Under Curve (AUC) (5) is used to measure the performance of classifiers over imbalanced datasets using the graph of the Receiver Operating Characteristic (ROC) [16]. In these graphics, the trade off between the benefits (TPrate) and cost(FPrate) can be visualized, and represent the fact that the capacity of any classifier cannot increase the number of true positives without also increasing the false positives. AUC summarizes the performance of the learning algorithm in a number.

$$AUC = \frac{1 + TPrate - FPrate}{2} \tag{5}$$

where:

$TPrate$: Correctly classified positive cases that belong to the positive class

$FPrate$: Negative cases that were misclassified as positive examples

3.1. Experimental study

The AUC result values is studied with this already balanced dataset. Table 2 shows that the AUC results of the data-balancing algorithm applying the Covariance Matrix with its CovI and CovO variants are similar or comparable with respect to the state-of-the-art oversampling algorithms, using as C4 .5 classifier.

Table 2. AUC of the data balancing algorithms with generation of oversampling classes of the state-of-the-art, CovI and CovO

Algorithms	Iris	glass1	Pima	vehicle3	glass-0-1-2-3_ vs _4-5-6	ecoli2	new thyroid2
ADASYN	**1**	0.74	0.73	0.74	0.88	0.91	**0.98**
Borderline–SMOTE	0.99	**0.77**	0.70	0.65	0.82	0.89	0.95
SMOTE–ENN	0.99	0.74	0.74	0.71	**0.93**	0.89	0.92
SMOTE–RSB	0.97	0.72	**0.75**	0.73	0.90	0.89	0.96
SMOTE–TL	0.99	0.74	0.72	**0.79**	0.90	0.89	0.93
SMOTE	**1**	**0.77**	0.74	0.72	0.84	**0.92**	0.92
SPIDER	0.99	0.74	0.72	0.71	0.92	0.89	0.95
Original	**1**	0.72	**0.75**	0.72	0.90	0.85	0.96
SMOTE–CovO	**1**	0.71	0.72	0.71	0.92	0.86	0.95
SMOTE–CovI	0.95	0.72	0.70	0.72	0.86	0.86	0.96

For the statistical analysis of the results, hypothesis-testing techniques were used [18, 19]. In both experiments, the Friedman and Iman-Davenport tests were used [23], in order to detect statistically significant differences between groups of results. The Holms test was also carried out [24], with the aim of finding significantly higher algorithms. These tests are suggested in the studies presented in [18, 19, 23], where it is stated that the use of these tests is highly recommended for the validation of results in the field of automated learning. Table 3 shows the ranking obtained by the Friedman test for the experiment. Although the algorithm with the best ranking was ADASYN, Holm's test performed below will demonstrate to what extent this algorithm can be significantly superior to the one proposed in the research.

Table 3. Friedman's Test

Algorithms	Ranking
ADASYN	3.4286
Borderline–SMOTE	6.9286
SMOTE–ENN	5.4286
SMOTE–RSB	4.9286
SMOTE–TL	5.2857
SMOTE	4.5714
SPIDER	5.6429
Original	5
SMOTE–CovO	6.3571
SMOTE–CovI	7.4286

Table 4. Holms test with $\alpha = 0.05$, taking ADASYN as a control method

i	Algorithms	$Z = \frac{(R_o - R_i)}{SE}$	p–value	Holm	Hypothesis
9	SMOTE–CovI	2.47	0.01	0.005	Reject
8	Borderline–SMOTE	2.16	0.03	0.006	Reject
7	SMOTE–CovO	1.80	0.07	0.007	Accept
6	SPIDER	1.36	0.17	0.008	Accept
5	SMOTE–ENN	1.23	0.21	0.01	Accept
4	SMOTE–TL	1.14	0.25	0.012	Accept
3	Original	0.97	0.33	0.01	Accept
2	SMOTE–RSB	0.92	0.35	0.02	Accept
1	SMOTE	0.70	0.48	0.05	Accept

Table 4 summarizes the results of Holms test taking ADASYN as a control method, all hypotheses with p–value ≤ 0.05 are rejected; showing that ADASYN is significantly superior to the SMOTE-CovI and Borderline-SMOTE algorithms. In the case of SMOTE-CovO, SPIDER, SMOTE_ENN, SMOTE_TL, Original, SMOTE-RSB and SMOTE, the null hypothesis is accepted, this mean that there are not significant differences between ADASYN and them, so it can be concluded that they are as effective.

On the other hands, the results achieved on Nemenyi's post hoc comparisons for $\alpha = 0.05$ and adjusted p-values are shown in:

Table 5. Nemenyi's test with P-values $\alpha = 0.05$

i	Algorithms	$Z = \frac{(R_0 - R_i)}{SE}$	p–value	Hypothesis
45	ADASYN vs. SMOTE–CovI	2.471658	0.013449	Reject
43	ADASYN vs. SMOTE–CovO	1.809606	0.070357	Accept
42	SMOTE vs. SMOTE–CovI	1.76547	0.077485	Accept
41	SMOTE-RSB vs. SMOTE–CovI	1.544786	0.122398	Accept
40	Original vs.SMOTE–CovI	1.500649	0.133446	Accept
37	SMOTE-TL vs. SMOTE–CovI	1.324102	0.185469	Accept
34	SMOTE-ENN vs. SMOTE–CovI	1.235829	0.216522	Accept
31	SMOTE vs. SMOTE–CovO	1.103419	0.269845	Accept
30	SPIDER vs. SMOTE–CovI	1.103419	0.269845	Accept
25	SMOTE-RSB-Is0 vs. SMOTE–CovO	0.882735	0.37738	Accept
24	Original vs. SMOTE–CovO	0.838598	0.401695	Accept
20	SMOTE–CovO vs. SMOTE– CovI	0.662051	0.507938	Accept
19	SMOTE-TL vs. SMOTE–CovO	0.662051	0.507938	Accept
18	SMOTE-ENN vs. SMOTE–CovO	0.573778	0.566118	Accept
15	SPIDER vs. SMOTE–CovO	0.441367	0.658947	Accept
12	Borderline-SMOTE vs. SMOTE–CovO	0.353094	0.724018	Accept
10	Borderline-SMOTE vs. SMOTE–CovI	0.308957	0.757354	Accept

Nemenyi's procedure rejects those hypotheses that have an adjusted p-value ≤ 0.001111.

Nemenyi test is a test that intended to find difference on the groups of data after a statistical test of multiple comparisons. If it has rejected the null hypothesis that the performance of the comparisons on the groups of data then is similar. The test does pair-wise tests of performance. As can be observed, ADASYN and SMOTE-CovI has a similar performance, while the rest of pair-wise has a different performance.

4. Conclusions and future work

In this paper, a new algorithm is proposed to generate synthetic instances of the minority class, using the Covariance Matrix. The experimental study carried shows the effectiveness of the proposed algorithm compared to eight recognized algorithms of the state-of-the-art. SMOTE-Cov showed similar or comparable results, taking into account the results of the AUC curve of the C4.5 classifier and using non-parametric tests to demonstrate that there are no significant differences between them, with the exception of the ADASYN versus the SMOTE-CovI variant. This can be influenced because the attributes present in the studied datasets come from other intervals and not from the actual attribute within the dataset.

Having results comparable to those of the state-of-the-art, for these datasets, allows in the future to extend the experimentation to datasets with tens, hundreds or thousands of attributes and with strong dependency relationships. It is also intended to use covariance regularization (Shrinkage) to balance data, where the number of positive instances is less than the number of attributes. The last recommendation is study the extension of the proposed algorithm to multi-class classification problems.

5. Acknowledgments

We would like to thanks VLIR (Vlaamse Inter Universitaire Raad, Flemish Interuniversity Council, Belgium) for supporting this work under the Networks 2019 Phase 2 Cuba ICT and the Play! for food: Improving food production and social welfare in the province of Camagüey project.

References

[1] Maher Maalouf and Theodore B Trafalis. Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, 55(1):168–183, 2011.

[2] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[3] Khudran Alzhrani, Ethan M Rudd, C Edward Chow, and Terrance E Boult. Automated big security text pruning and classification. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3629–3637. IEEE, 2016.

[4] Alberto Fernández, Salvador García, María José del Jesus, and Francisco Herrera. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, 159(18):2378–2398, 2008.

[5] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

[6] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[8] Yueh-Min Huang, Chun-Min Hung, and Hewijin Christine Jiau. Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, 7(4):720–747, 2006.

[9] Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

[10] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 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):463–484, 2011.

[11] Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*, 2016.

[12] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in

imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[13] Jerzy Stefanowski and Szymon Wilk. Selective preprocessing of imbalanced data for improving classification performance. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 283–292. Springer, 2008.

[14] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[15] Robert M Sanner and J-JE Slotine. Gaussian networks for direct adaptive control. *IEEE Transactions on neural networks*, 3(6):837–863, 1992.

[16] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[17] UCI Machine Learning Repository, `https://archive.ics.uci.edu/ml/index.php`. Last accessed 14 Feb 2019

[18] Salvador García and Francisco Herrera. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary computation*, 17(3):275–306, 2009.

[19] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

[20] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[21] KEEL-dataset repository, `http://www.keel.es/`. Last accessed 14 Feb 2019

[22] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[23] Ronald L Iman and James M Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods*, 9(6):571–595, 1980.

[24] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[25] Yoan Martínez-López, Julio Madera, Ansel Y. Rodríguez-González and Stephen Barigye Cellular Estimation Gaussian Algorithm for Continuous Domain *Journal of Intelligent & Fuzzy Systems*, pages 1–11 Elsevier, 2019.

[26] Yoan Martínez-López, Ansel Y. Rodríguez-González and Stephen Barigye. CUMDANCauchy-C1: a cellular EDA designed to solve the energy resource management problem under uncertainty. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*,(ACM), pages 13–14, 2019

[27] E. Ramentol, I. Gondres, S. Lajes, R. Bello, Y. Caballero, C. Cornelis and F. Herrera. Fuzzy-rough imbalanced learning for the diagnosis of High Voltage Circuit Breaker maintenance: The SMOTE-FRST-2T algorithm. *Engineering Applications of Artificial Intelligence*, pages 134–139 Elsevier, 2016.

[28] Yuka Takahashi. Yu Fujimoto and Yasuhiro Hayashi Forecast of infrequent wind power ramps based on data sampling strategy. *Energy Procedia*, pages 496—503 Elsevier, 2017.

[29] Christian Promper, Engel Dominik , and Robert C. Green Anomaly detection in smart grids with imbalanced data methods. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8 IEEE, 2017.

[30] Ireimis Leguen-deVarona, Julio Madera, Yoan Mart´nez-López and José Carlos Hernández-Nieto SMOTE-Cov: A new over-sampling method based on the Covariance Matrix. In *3rd EAI International Conference on Computer Science and Engineering and Health Serices (COMPSE 2019)*, pages 0–0 0 COMPSE, 2019.

[31] Michal Koziarski Radial-Based Undersampling data classification. Pattern Recognition, 102, pages 107262 Elsevier,2020

[32] Zhu Tuanfei, Lin Yaping, and Liu Yonghe Improving interpolation based oversampling for imbalanced data learning Knowledge Based System, 187, pages 104826 Elsevier,2020

[33] Enislay Ramentol, Yaile Caballero, Rafael Bello, Francisco Herrera SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl Inf Syst 33, 245–265 (2012). https://doi.org/10.1007/s10115-011-0465-6

[34] Robertas Damasevicius Analysis of binary feature mapping rules for prometer recognition in imbalanced DNA sequence datasets using support vector machine In *4th International IEEE Conference Intelligent Systems* IEEE,2008