Collective of base classifiers for mining imbalanced data

Joanna Jedrzejowicz[1][0000−0003−4979−5476]
and Piotr Jedrzejowicz[2][0000−0001−6104−1381]

[1] Institute of Informatics, Faculty of Mathematics, Physics and Informatics, University of Gdansk, 80-308 Gdansk, Poland, joanna.jedrzejowicz@ug.edu.pl
[2] Department of Information Systems, Gdynia Maritime University, 81-225 Gdynia, Poland, p.jedrzejowicz@umg.edu.pl

Abstract. Mining imbalanced datasets is a challenging and difficult problem. In this paper we adress it by proposing GEP-NB classifier based on the oversampling technique. It combines two learning methods – Gene Expression Programming and Naïve Bayes, which cooperate to produce a final prediction. At the pre-processing stage a simple mechanism for generating synthetic minority class examples and balancing the training set is used. Next, two genes g1 and g2 are evolved using Gene Expression Programming. They differ by applying in each case a different procedure for selecting synthetic minority class examples. If the class prediction by g1 agrees with the class prediction made by g2, their decision is final. Otherwise the final predictive decision is taken by the Naïve Bayes classifier. The approach is validated in an extensive computational experiment. Results produced by GEP-NB are compared with performance of several state-of-the-art classifiers. Comparisons show that GEP-NB offers a competitive performance.

Keywords: Imbalanced data · Oversampling · Gene expression programming.

1 Introduction

Datasets with an unequal distribution of classes are commonly referred to as imbalanced ones. Unequal distribution of classes is encountered in numerous real-life situations such as, for example, fault diagnosis, medical diagnosis, fraud detection, credit rating, and many other critical applications. During the last two decades numerous approaches, techniques and algorithms have been proposed to deal with mining imbalanced datasets. Our research goal is to extend the range of available approaches for mining imbalanced datasets by proposing and validating an effective new classifier based on a novel oversampling procedure and GEP and NB learners integrated using a semi-ensemble architecture. In such an architecture at least two out of three base learners have to agree when taking the predictive decision.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

Algorithms and methods proposed for mining imbalanced datasets can be broadly categorized as data-level, algorithm-level, and hybrid approaches. Data-level approaches can be further divided into oversampling and undersampling methods. Their goal is to transform the dataset used for learning prior to applying some learners. Such transformation usually leads to achieving the balanced or, at least, a better balanced distribution of classes.

Our motivation for using GEP and NB for inducing base classifiers has been based on the earlier performance of both techniques in data mining applications. A review of numerous successful GEP applications in machine learning can be found in [13]. Naïve Bayes is a probabilistic classifier that can achieve a high accuracy level [10]. Besides, NB learners are scalable and require some parameters linear in the number of variables in a learning problem. Both types of learners, that is GEP and NB are based on different philosophies, which makes their prediction fairly independent. The above feature plays an important role in the proposed approach where NB has a decisive role in the case when GEP induced base learners produce different predictions.

The rest of the paper is organized as follows. Section 2 contains a concise overview of the related work. Section 3 presents the proposed learner named GEP-NB. Section 4 discusses results of an extensive computational experiment carried out to validate the approach. Final Section 5 contains conclusions and suggestions for future research.

2 Related work

In this Section we will briefly review several learners, including those used for mining imbalanced datasets, that are used for comparison and validation purposes in Section 4.

The simplest approach to balancing imbalanced datasets are random undersampling (RUS) and random oversampling (ROS). RUS works through random elimination of instances from majority class, and ROS through random replication of minority class instances. Both approaches have disadvantages – RUS may eliminate potentially informative examples and ROS may cause an overfitting.

One of the most often used approaches for mining imbalanced datatests is SMOTE - an oversampling technique proposed in [3]. In SMOTE the minority class is oversampled by introducing synthetic instances selected randomly and iteratively along the line segments joining some of the k minority class nearest neighbors until the balance between classes is achieved. Well known extension of SMOTE is the ADASYN method [11]. In [24] an approach named LLE for enhancing the SMOTE by incorporating the locally linear embedding algorithm was proposed. Another improvement of SMOTE obtained by introducing the PCA framework was proposed in [20]. Further, numerous, extensions and modification of SMOTE are reviewed in [6].

An approach to oversampling strategy using a rough–granular computing approach (RGA) was proposed in [2]. Another approach based on the rough set theory was proposed in [5]. The authors proposed a method for feature selection

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

for imbalanced datasets using the neighborhood rough set theory. The approach assumes that imbalanced distribution of classes reflects the definition of the feature significance. A discernibility-matrix-based feature selection method is next defined and used in the feature selection algorithm (RSFSAID). Finally, a particle swarm optimization algorithm is suggested to optimize parameters of the algorithm.

Recently, an approach for enhancing the performance of oversampling methods for class imbalance classification was proposed in [16]. The authors propose a novel hybrid technique named ant colony optimization resampling (ACOR) to overcome class imbalance.

In [27] the authors claim that oversampling methods are often disrupted by noise when data are not well separated. As a remedy they propose the framework using the Laplacian eigenmaps (EIGEN FRAMEWORK) to find an optimal dimensional space, where the data are well separated and the generation of noise by SMOTE based oversampling methods can be avoided or minimized.

Tomek Link (TL) is an undersampling technique originating from [21]. One of the oldest approaches to undersampling is the Edited Nearest Neighbors (ENN) algorithm based on Wilsons rules [26]. The default behavior of ENN is to remove examples from the majority class that are misclassified by their k nearest neighbors. The Repeated ENN (RENN) runs the ENN algorithm repeatedly until all instances remaining have a majority of their neighbors with the same class [25]. One Side Selection (OSS) algorithm proposed in [14] is another undersampling technique. The algorithm starts with constructing a 1-NN classifier from dataset containing all minority class instances and a single, randomly drawn, majority class instance. Next, it appends misclassified instances from the set of remaining ones and removes borderline and noisy instances using Tomek links. An improvement of the OSS was proposed in [15]. The proposed Neighbouring Cleaning Rule (NCR) algorithm is similar to OSS, except that to identify uninformative and noisy data the edited nearest-neighbor rule is used instead of the TL.

Undersampling approach for learning Naïve Bayes classifiers for mining imbalanced datasets (NBU) was presented in [1].

In the recent years several undersampling algorithms using clustering have been proposed. One of the first was the algorithm Fast-CBUS proposed by [19]. The idea was to group majority instances from the training set into clusters. A separate classifier is then trained for each cluster. An unlabeled instance is classified as the majority class if it does not fit into any of the clusters. Otherwise, separate classifiers induced earlier on are used to return the classification results, and the results are weighted by the inverse-distance from the clusters.

Well performing Clustering-based undersampling (CBU) was proposed in [17]. The authors introduce two undersampling strategies aiming at reducing the number of instances in the majority class to balance the training dataset. The idea is to partition majority class instances into clusters. The number of clusters is set to the number of instances in the minority class. The first strategy is to use cluster centroids as the majority class representation, while the second

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

strategy uses their nearest neighbors instead. During the learning phase the AdaBoost with C4.5 ensemble classifier was induced.

Combining a clustering-based undersampling based with instance selection was the idea of [22]. The cluster based instance selection (CBIS) uses two components. The first, groups instances from the majority class into clusters, and the second filters out unrepresentative ones from each cluster. For clustering the affinity propagation (AP) algorithm proposed in [9] is used and for instance selection, either a genetic algorithm, or IB3, or DROP3 can be used (for instance selection algorithms see [25]).

An effective approach to mining imbalanced datasets is using the ensemble learning techniques. The idea is to combine several base-learners into ensembles of classifiers. One of the first was the ensemble oversampling algorithm named SMOTEBoost, proposed in [4].

Combining undersampling and oversampling techniques in an ensemble learner was proposed in [23]. The learner known as UnderBagging and OverBagging (UOBag) works as follows: In UnderBagging, several subsets of instances are created by undersampling majority class randomly to construct classifiers. In a similar way, OverBagging creates subsets of instances by oversampling minority classes randomly. When a new instance arrives the majority vote decides on class prediction.

Ensemble classifier for imbalanced data based on feature space partitioning and hybrid metaheuristics (AdaSSGACE) was proposed in [18].

3 GEP-NB classifier

3.1 General idea of the GEP-NB

The proposed GEP-NB classifier is based on the oversampling technique. It combines two learning methods – Gene Expression Programming and Naïve Bayes classifier which are used to produce a collective learner responsible for the final prediction. GEP-NB can be used for solving binary classification problems. At the pre-processing stage a simple mechanism for generating a synthetic minority class examples and balancing the training set is used. For balancing purposes, available minority class examples from the training set are supplemented by some synthetic minority class examples to produce an expanded minority class (EMC) training dataset consisting of original plus synthetic examples. During the first phase of constructing the EMC dataset, original minority examples are randomly replicated and attached to the current EMC. The number of minority class instances drawn in such a way, denoted as Ms, is set by the user. Each instance from the minority class can be drawn many times. The number of replicated minority instances plus the original set of minority instances in the training set should exceed the number of majority instances in the training set. The size of the EMC is controlled by the parameter Ms. All replicated minority instances are subject to mutation.

For each replicated instance, the mutation procedure starts with randomly selecting (based on the uniform distribution) subset of features (without class

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

labels), which will undergo a mutation. Selected features are modified according to the following heuristic rules: Boolean values are reversed, integer values are changed by randomly modifying (adding or subtracting) x percent of their value, and taking the integer part of the result, real values are changed by randomly adding or subtracting x percent of their value, where x is a parameter, set by the user. The idea is to produce some synthetic minority class instances constructed from the original minority class samples using the proposed random mutation procedure. The mutation scale is controlled by the parameter x, and the quality of thus produced synthetic samples is controlled by the subsequent selection procedures used for achieving a balance between minority and majority classes. In numerous applications including evolutionary computations, genetic programming, and population-based meta-heuristics, mutation procedures are used as means for improving diversification of solutions, assuring better convergence, and helping to escape from local optima. In our case, the role of mutation is to diversify synthetic samples and still keeping them somehow similar to original minority instances. To avoid the negative influence of the outliers and to keep samples fairly uniformly distributed we propose two specialized selection procedures. Our approach has been inspired by population-based techniques that have proven effective for solving a variety of difficult problems.

Next, two genes $g1$ and $g2$ are evolved using Gene Expression Programming. Both learners have the form of expression trees induced under the criterion of geometric mean (G) value which should be maximized. The choice of geometric mean as the main criterion is motivated by the fact that G is one of the most often used metrics for evaluating the performance of learners designed for mining imbalanced datasets. Besides, the value of G is closely correlated with values of other metrics commonly used in the case of imbalanced datasets mining. The learners $g1$ and $g2$ differ by applying in each case a dedicated selection procedure for reducing the EMC to balance minority and majority datasets:

- In the case of $g1$, the centroid of the EMC is identified, and the Euclidean distance between the centroid and each of the instances in the EMC is calculated. At this point we apply an instance reduction procedure to obtain the reduced EMC with fairly uniform distribution of instances in the solution space as shown in Fig. 2. Reduction aims at balancing majority and minority classes.
- In the case of $g2$, the centroid of the majority class is identified, and the distance between this centroid and each of the instances in the EMC is calculated. At this point we apply an instance reduction procedure to obtain the reduced EMC by discarding instances that are close to the centroid of the majority class as shown in Fig. 3, until majority and minority classes become balanced.

At the learning stage, classifiers $g1$ and $g2$, and a Naïve Bayes classifier play the role of base learners. Naïve Bayes learner is induced using a subset of the training set involving instances for which predictions produced by $g1$ and $g2$ have differed. The above learners are expected to maximize the value of the respective geometric mean. Classifiers produced by GEP have the form of expression trees.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

When classifying instances belonging to the test set, if for an instance the class prediction by $g1$ agrees with that of $g2$, their decision is final. Otherwise, the final predictive decision is made by the Naïve Bayes classifier. An example of an expression tree, a formal description of the approach and its computational complexity analysis is given in the next subsections.

3.2 Formal description of the approach

Gene Expression Programming (GEP), introduced by Ferreira [8] is a meta-heuristic which can be used in several areas, classification included. It combines the idea of genetic algorithms and genetic programming and makes use of a population of genes. Each gene is a linear structure divided in two parts. The first part, head, contains functions and terminals while the second part, tail, contains only terminals. For this study terminals are of type $(oper, attr, const)$, where the value of $const$ is in the range of attribute $attr$ and $oper$ is a relational operator from $\{<, \leq, >, \geq, =, \neq\}$. Functions are from the set $\{AND, OR, NOT, XOR, NOR\}$. For a fixed instance x from the dataset, the value $g(x)$ of a gene g is boolean and thus a gene can be treated as a binary classifier.

Learning the best gene classifier is an iterative process which starts with a random population of genes. In each iteration the population is subjected to operations such as: mutation, root transposition, transposition of insertion sequence, 1-point and 2-point recombination. Each operation is performed with a probability which is a parameter of the process. More details on applying GEP can be found in [12].

Considering our hybrid classifier, the first step is to oversample the minority class by mutating random rows, as described in Fig. 1. The mutation of an

Require: data from minority class MinC, parameter Ms
Ensure: expanded minority class EMC of size Ms.
1: EMC=$\emptyset$
2: **for** $i = 0$ to Ms **do**
3: draw random data row rw from MinC
4: draw random subset AT of attributes
5: **for all** $at \in AT$ **do**
6: mutate $rw(at)$ to $r\bar{w}(at)$ applying (1)
7: **end for**
8: add $r\bar{w}$ to EMC
9: **end for**
10: **return** EMC

Fig. 1. Oversampling to generate expanded minority class

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

attribute is defined as:

$$r\bar{w}(at) = \begin{cases} 1 - at & \text{if } at \text{ boolean} \\ (int)(rand(at * (+/-(1+x))) & \text{if } at \text{ integer} \\ rand(at * (+/-(1+x))) & \text{otherwise} \end{cases} \quad (1)$$

In the next step one of two different selection procedures is applied to EMC to balance majority and minority sets. The procedures are given in Fig. 2 and Fig. 3, respectively. Finally, the algorithm shown in Fig. 4 is applied to learn the classifier, test it and calculate entries in confusion matrix and the respective performance measures.

Require: expanded minority class EMC , data from majority class MajC
Ensure: balanced dataset
1: calculate centroid CN of MajC
2: **for all** $x \in$EMC **do**
3: calculate $dist(x, CN)$
4: **end for**
5: discard from EMC instances closest to CN to balance with MajC
6: **return** EMC $\cup$ MajC

Fig. 2. Instance reduction with equal distribution

Require: expanded minority class EMC, data from majority class MajC
Ensure: balanced data set.
1: calculate centroid CN of EMC
2: define quartiles for $\{dist(x, CN)\ x \in$ EMC$\}$
3: **for** $i = 1, 2, 3, 4$ **do**
4: let n_i=number of elements in quartile i
5: $Q_i = n_i/(n_1 + n_2 + n_3 + n4)$
6: **end for**
7: **repeat**
8: t=random
9: let $(t > Q_{i-1} \wedge (t \leq Q_i))$
10: discard random instance from quartile i
11: **until** EMC and MajC are balanced
12: **return** EMC $\cup$ MajC

Fig. 3. Instance reduction with centroids

As far as computational complexity is concerned, for Fig. 1 it is $O(|EMC|)$, where EMC is the expanded minority class. For Fig. 2 and Fig. 3 it is $O(|EMC|^2)$, and finally, for Fig. 4 it is bounded by the complexity of 2 and the complexity

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

Require: minority class MinC, majority class MajC, testing data Test
Ensure: measures of classification quality
1: use algorithm from Fig.1 to expand minority class MinC to extended minority class EMC
2: use algorithm from Fig. 2 to generate balanced training set $Train1$
3: generate best possible gene $g1$ with $Train1$
4: use algorithm from Fig. 3 to generate balanced training set $Train2$
5: generate best possible gene $g2$ with $Train2$
6: **for all** $(x, c) \in$ Test **do**
7: calculate $g1(x) = v1$ and $g2(x) = v2$
8: **if** $v1 = v2$ **then**
9: $class = v1$
10: **else**
11: apply Naïve Bayes classifier to x to define $class$
12: **end if**
13: compare $class$ with c and modify TP, FN, FP, TN respectively
14: **end for**
15: calculate quality measures from TP, TN, FP, FN
16: **return** quality measures

Fig. 4. Learning best genes and testing

of learning the best gene which is $O(nIt \times popSize \times |dataset|)$, where nIt is the number of iterations in GEP, $popSize$ is the population size and $|dataset|$ is the size of the dataset.

4 Computational experiment

To validate the proposed approach we have carried out an extensive computational experiment. It has covered 100 of the imbalanced datasets available in KEEL Dataset Repository (https://sci2s.ugr.es/keel/imbalanced.php). Datasets originate from [6] and [7]. Full information about the above datasets including dataset names, number of instances, number of features and value of the imbalanced ratio can be found in the KEEL Dataset Repository.

In the experiment we have applied 5-folds cross validation procedure which has been repeated 10 times. All the reported values are averages from the above scheme. Geometric Mean (G) and Area Under the ROC Curve (AUC) were selected as performance metrics.

GEP-NB has been run with the following settings, identical for all considered datasets:

- GEP population size: 100
- Number of iterations in GEP: 100
- Selection rule: tournament from the pair
- Mutation probability: pm = 0.5
- Root Insertion Sequence Transposition probability: pris = 0.2

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

- Insertion Sequence probability: is = 0.2
- 1-point recombination probability: pr1 = 0.2
- 2-points recombination probability: pr2 = 0.2
- Size of the EMC: twice the number of majority instances in the training set
- The value of $x = 5\%$.

5 Comparative analysis

To evaluate the proposed approach several comparisons with the state-of-the-art algorithms for mining imbalanced datasets have been carried out. Table 1 shows average geometric mean (G) obtained using the following learners: Naïve Bayes Undersampling (NBU), Edited Nearest Neighbors (ENN), Neighbouring Cleaning Rule (NCR), One Side Selection (OSS), Repeated Edited Nearest Neighbors (RENN), Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), Tomek Links (TL), the proposed oversampling scheme denoted Population-based Oversampling (PBO), and GEP-NB proposed in this paper. In all cases values of the respective metric have been calculated as an average from several runs of the 5-cross-validation scheme. Results for NBU, ENN, NCR, OSS, RENN, RUS, SMOTE and TL are taken from [1]. In all of the above cases, as well as in the case of PBO, final results were obtained using Naïve Bayes classifier. To determine whether there are any significant differences among results from Table 1, produced by different classifiers we used the Friedman ANOVA by ranks test. The null hypothesis state that there are no such differences. With Friedman statistics equal to $75,39$ and p-value equal to 0.00000 the null hypothesis should be rejected at the significance level of 0.05. However, the Kendall concordance coefficient expressing the simultaneous association (relatedness) between the considered samples, with the value of 0.1948 tells that there is a limited degree of relatedness between the considered samples.

Another comparison involved GEP-NB and the following ensemble learners: Random Undersampling with Boosting (RUSBoost), ensemble classifier based on feature space partitioning with hybrid metaheuristics (AdaSSGACE) using KNN and SVM base classifiers, and Underbagging and Overbagging Ensemble Learner (UOBag). The performance metric used in the comparison is the area under the ROC curve (AUC). In all cases values of the respective metric have been calculated as an average from several runs of the 5-cross-validation scheme. Results for RUSBoost, AdaSSGACE and UOBag are taken from [18]. The respective results are shown in Table 2. Table 3 shows results obtained by GEP-NB, clustering based undersampling (CBU) proposed in [17] and two versions of the ensemble classifier proposed in [22], and known as undersampling by combining clustering analysis and instance selection (CBIS). Both version of this algorithm use clustering by passing messages between data points technique [9] and IB3 instance selection algorithm [25]. First version uses boosting for constructing ensembles, and the second one uses bagging for the same purpose. All results have been obtained using the 5-CV scheme. Results for CBU and CBIS are taken from the original sources. In all cases the performance metric is the area under the ROC

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

Table 1. Comparison of the average geometric mean (G) values obtained by the analyzed classifiers.

Dataset	NBU	ENN	NCR	0SS	RENN	RUS	SMOTE	TL	PBO	GEP-NB
abalone19	0,659	0,695	0,695	0,695	0,695	0,708	0,679	0,694	0,713	**0,746**
dermatology6	0,988	0,966	0,966	0,966	0,879	0,966	0,975	0,966	0,997	**0,999**
ecoli-0-1-4-6-VS-5	0,880	0,858	0,858	0,853	0,858	0,861	0,862	0,858	0,934	**0,949**
ecoli-0-1-4-7-VS-2-3-5-6	**0,960**	0,922	0,924	0,926	0,915	0,908	0,927	0,925	0,860	0,875
ecoli-0-1-4-7-VS-5-6	**0,958**	0,948	0,948	0,945	0,948	0,826	0,943	0,949	0,953	0,928
ecoli-0-2-3-4-vs-5	0,900	0,856	0,858	0,869	0,856	0,816	0,874	0,863	0,912	**0,957**
ecoli-0-3-4-6-vs-5	0,922	0,849	0,850	0,853	0,849	0,766	0,854	0,849	0,879	**0,951**
ecoli-0-3-4-7-vs-5-6	**0,941**	0,925	0,925	0,919	0,925	0,928	0,906	0,925	0,911	0,920
ecoli-0-3-4-vs-5	0,894	0,835	0,844	0,860	0,835	0,750	0,860	0,844	0,908	**0,927**
ecoli-0-4-6-vs-5	0,872	0,846	0,846	0,853	0,846	0,848	0,857	0,846	0,917	**0,949**
ecoli-0-6-7-vs-5	0,915	0,874	0,879	0,871	0,874	0,838	0,883	0,883	0,914	**0,950**
ecoli1	0,885	0,850	0,852	0,839	0,853	0,872	0,842	0,848	0,892	**0,931**
ecoli2	**0,941**	0,928	0,929	0,928	0,927	0,921	0,942	0,928	0,845	0,939
ecoli3	0,894	0,890	0,892	0,920	0,873	0,923	0,908	0,907	0,857	**0,924**
ecoli4	0,950	0,916	0,916	0,920	0,916	0,916	0,934	0,917	0,879	**0,982**
glass-0-1-4-6-vs-2	0,720	0,685	0,690	0,680	0,699	0,618	0,703	0,696	0,702	**0,737**
glass0	0,826	0,771	0,800	0,810	0,755	0,800	0,793	0,804	0,799	**0,875**
glass1	0,660	0,663	0,662	0,669	0,702	0,702	0,637	0,656	0,678	**0,761**
glass4	0,830	0,723	0,726	0,714	0,723	0,760	0,752	0,728	0,766	**0,969**
glass6	0,864	0,854	0,854	0,825	0,888	0,855	0,827	0,856	0,869	**0,971**
haberman	0,656	0,670	0,676	0,658	0,671	0,611	0,645	0,657	0,712	**0,731**
iris0	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**
new-throid1	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	0,945	0,974
new-thyroid2	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	0,966	0,978
page-blocks-1-3-vs-4	0,992	0,902	0,908	0,908	0,902	0,902	0,909	0,909	0,913	**0,999**
page-blocks0	**0,954**	0,936	0,935	0,928	0,937	0,935	0,932	0,932	0,905	0,893
pima	**0,815**	0,799	0,809	**0,815**	0,789	0,814	0,815	0,813	0,814	0,748
poker-8-vs-6	0,531	0,437	0,439	0,427	0,437	0,456	0,500	0,577	0,588	**0,801**
segment0	**0,987**	0,982	0,982	0,982	0,982	0,982	0,980	0,982	0,981	0,984
vehicle0	**0,904**	0,806	0,805	0,823	0,801	0,809	0,820	0,811	0,834	0,901
vehicle1	0,740	0,709	0,713	0,717	0,708	0,719	0,717	0,713	0,726	**0,754**
vehicle2	0,920	0,861	0,859	0,850	0,857	0,834	0,849	0,857	0,789	**0,926**
vehicle3	0,762	0,701	0,700	0,698	0,700	0,697	0,699	0,698	0,798	**0,926**
winequality-red-4	**0,694**	0,659	0,653	0,651	0,660	0,626	0,653	0,650	0,616	0,679
winequality-red-8-vs-6-7	0,695	0,713	0,717	0,669	0,721	0,674	0,651	0,713	0,674	**0,815**
wisconsin	0,978	0,975	0,977	**0,993**	0,974	0,983	0,983	0,982	0,979	0,973
yeast-0-2-5-6-vs-3-7-8-9	0,816	0,761	0,762	0,762	0,764	0,742	0,755	0,760	0,776	**0,848**
yeast-0-2-5-7-9-vs-3-6-8	0,932	0,916	0,916	0,888	0,916	0,907	0,816	0,915	0,923	**0,939**
yeast-0-3-5-9-vs-7-8	0,721	0,695	0,690	0,712	0,680	0,713	0,731	0,698	0,698	**0,727**
yeast-1-vs-7	0,801	0,802	0,800	0,800	**0,805**	0,792	0,776	0,800	0,703	0,737
yeast-2-vs-4	0,864	0,833	0,835	0,833	0,831	0,822	0,834	0,839	0,837	**0,956**
yeast-2-vs-8	0,838	0,835	0,836	0,828	0,835	**0,850**	0,793	0,836	0,818	0,833
yeast5	**0,989**	0,987	0,986	0,986	0,987	0,976	0,982	0,986	0,935	0,978
Average	0,862	0,833	0,835	0,834	0,832	0,824	0,832	0,835	0,840	**0,892**

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

Table 2. Comparison of the average geometric mean (G) values obtained by the analyzed classifiers.

Dataset	GEP-NB	RUSBoost	AdaSSGACEKNN.40	AdaSSGACESVM.40	UOBag
ecoli1	**0,933**	0,884	0,890	0,872	0,876
ecoli3	**0,927**	0,840	0,864	0,785	0,886
iris0	**1,000**	0,990	0,999	0,841	0,970
page-blocks0	0,917	**0,956**	0,931	0,751	0,953
pima	**0,751**	0,725	0,738	0,589	0,730
vehicle1	0,757	**0,786**	0,778	0,651	0,745
yeast1	**0,722**	0,701	0,711	0,688	0,720
yeast3	**0,945**	0,919	0,897	0,891	0,919
glass1	**0,772**	0,780	0,750	0,639	0,739
glass6	**0,971**	0,921	0,903	0,641	0,901
glass-0-1-6_vs_2	**0,773**	0,700	0,708	0,611	0,629
ecoli4	**0,982**	0,896	0,940	0,907	0,867
glass-0-1-6_vs_5	0,960	0,954	0,867	0,733	**0,963**
glass5	0,966	0,949	0,804	0,675	**0,988**
dermatology6	**0,999**	0,966	0,966	0,749	0,938
shuttle-6_vs_2-3	**1,000**	0,902	0,965	0,843	0,948
poker-9_vs_7	**0,886**	0,590	0,740	0,636	0,556
yeast-2_vs_8	**0,840**	0,747	0,801	0,737	0,778
yeast4	**0,867**	0,827	0,799	0,509	0,763
led7digit-0-2-4-5-6-7-8-9_vs_1	**0,917**	0,894	0,856	0,785	0,881
ecoli-0-1-3-7_vs_2-6	**0,959**	0,896	0,848	0,588	0,867
winequality-red-8_vs_6	**0,827**	0,815	0,589	0,528	0,700
winequality-white-9_vs_4	**0,914**	0,893	0,645	0,576	0,714
yeast6	**0,914**	0,851	0,875	0,515	0,814
poker-8-9_vs_6	**0,835**	0,915	0,623	0,557	0,534
winequality-white-3-9_vs_5	**0,683**	0,674	0,561	0,531	0,576
shuttle-2_vs_5	**1,000**	**1,000**	0,986	0,672	**1,000**
winequality-red-3_vs_5	**0,848**	0,644	0,608	0,580	0,615
poker-8-9_vs_5	0,616	0,547	**0,631**	0,557	0,618
poker-8_vs_6	0,817	**0,915**	0,623	0,469	0,534
Average	**0,877**	0,836	0,797	0,670	0,791

Table 3. Comparison of the area under the ROC curve obtained by CBU, CBIS and GEP-NB classifiers

Dataset	CBU	CBIS		GEP-NB
		AP+IB3boost	AP+IB3boost	
Abalone9-18	0,831	0,849	**0,894**	0,808
Abalone19	0,728	0,624	0,617	**0,771**
Ecoli-0-vs-1	0,982	0,975	0,982	**0,993**
Ecoli-0-1-3-7-vs-2-6	0,804	0,877	0,879	**0,959**
Ecoli1	0,927	**0,958**	0,957	0,933
Glass0	0,873	**0,888**	0,885	0,876
Glass-0-1-2-3-vs-4-5-6	0,970	**0,980**	0,966	0,961
Glass-0-1-6-vs-2	**0,790**	0,775	0,713	0,773
Glass-0-1-6-vs-5	0,964	0,894	**0,987**	0,960
Glass1	0,824	0,812	**0,847**	0,772
Glass2	0,760	0,741	0,766	**0,822**
Glass4	0,853	0,944	**0,971**	0,970
Glass5	0,949	**0,994**	**0,994**	0,966
Glass6	0,905	0,951	0,934	**0,971**
Haberman	0,603	0,646	0,648	**0,737**
Iris0	0,990	0,990	0,990	**1,000**
New-thyroid1	0,973	0,979	**0,997**	0,975
New-thyroid2	0,924	0,976	**0,994**	0,978
Page-blocks0	0,986	**0,987**	**0,987**	0,917
Page-blocks-1-3-vs-2	0,992	0,998	0,997	**0,999**
Pima	0,758	0,771	**0,805**	0,751
Segment0	0,996	**0,999**	0,993	0,984
Shuttle-0-vs-4	**1,000**	**1,000**	**1,000**	**1,000**
Shuttle-2-vs-4	0,988	**1,000**	**1,000**	**1,000**
Yeast-1-2-8-9-vs7	0,692	**0,818**	0,775	0,661
Yeast-1-4-5-8-vs7	0,627	**0,777**	0,605	0,709
Yeast4	0,874	0,857	**0,914**	0,867
Yeast5	**0,987**	0,967	0,970	0,978
Yeast6	0,909	0,881	0,884	**0,914**
Average	0,878	0,893	0,895	**0,897**

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

curve (AUC). To determine whether there are any significant differences among results from Table 3, we again used the Friedman ANOVA by ranks test. The null hypothesis state that there are no such differences. With Friedman statistics equal to 7.301 and p-value equal to 0.0629 the null hypothesis should not be rejected at the significance level of 0.05. The Kendall concordance coefficient expressing the simultaneous association (relatedness) between the considered samples, with the value of 0.0811 tells that there is a limited degree of relatedness between the considered samples. It is worth noting that the above findings are not contradictory to the fact that the average performance of GEP-NB using the AUC metric and calculated over the sample of 29 datasets as shown in Table 3, is better than the performance of the remaining learners.

6 Conclusion

The paper contributes by proposing a new classifier for mining imbalanced datasets. The classifier named GEP-NB, has the following main features:

- At the preprocessing stage it uses an original oversampling method to balance minority and majority sets of instances. The approach is based on producing synthetic minority class instances by replication, mutation, and selection of instances from the original set of minority instances.
- At the learning stage a semi-ensemble strategy is used. It consists of developing two complex expression trees using the Gene Expression Programming paradigm. Both are supported by the Naïve Bayes learner used in the case when the class prediction from both genes differs.

An extensive computational experiment has shown that the performance of the proposed learner is competitive. Comparison with the state of the art single learners for mining imbalanced datasets proves that GEP-NB outperforms all of them. Comparison with the state of the art ensemble learners shows that GEP-NB either outperforms them or is at least as good as the best of the considered ensemble classifiers. The above findings allow to state that GEP-NB is a worthy addition to the family of classifiers dedicated to mining imbalanced dataset.

Competitive performance of GEP-NB can be attributed to a synergetic effect produced by three following factors – learners, oversampling procedure, and semi-ensemble architecture. Ensemble components including GEP and Naïve Bayes learners are themselves good performers, which have been confirmed by numerous studies. The proposed oversampling procedure has been constructed using the population-based paradigm where mutation of the population members and selection of better-fitted individuals play a vital role in the search for high-quality solutions. Finally, using the concept of semi ensemble learning with three base classifiers could mask some prediction errors. The main novelty and, at the same time, the solution of the research problem tackled by the paper is the integration of the above factors into a specialized learner producing good quality predictions when mining imbalanced datasets.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

Future research will focus on extending the approach, possibly by integrating oversampling and undersampling approaches for balancing minority and majority class instances with a view to increasing effectiveness of the approach. Another direction of studies could bring some improvements in the process of generating and selecting the synthetic minority instances.

References

1. Aridas, C.K., Karlos, S., Kanas, V.G., Fazakis, N., Kotsiantis, S.B.: Uncertainty based under-sampling for learning naive bayes classifiers under imbalanced data sets. IEEE Access **8**, 2122–2133 (2020)
2. Borowska, K., Stepaniuk, J.: A rough-granular approach to the imbalanced data classification problem. Appl. Soft Comput. **83** (2019)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
4. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: Lavrac, N., Gamberger, D., Blockeel, H., Todorovski, L. (eds.) Knowledge Discovery in Databases: PKDD 2003. Lecture Notes in Computer Science, vol. 2838, pp. 107–119. Springer (2003)
5. Chen, H., Li, T., Fan, X., Luo, C.: Feature selection for imbalanced data based on neighborhood rough sets. Inf. Sci. **483**, 1–20 (2019)
6. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer (2018)
7. Fernández, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. Int. J. Approx. Reason. **50**(3), 561–577 (2009)
8. Ferreira, C.: Gene expression programming: A new adaptive algorithm for solving problems. Complex Systems **13**(2) (2001)
9. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(8), 972–976 (2007)
10. Hand, D.J., Yu, K.: Idiot's bayes: Not so stupid after all? International Statistical Review / Revue Internationale de Statistique **69**(3), 385–398 (2001), http://www.jstor.org/stable/1403452
11. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IJCNN 2008. pp. 1322–1328 (2008)
12. Jedrzejowicz, J., Jedrzejowicz, P.: Experimental evaluation of two new gep-based ensemble classifiers. Expert Syst. Appl. **38**(9), 10932–10939 (2011)
13. Jedrzejowicz, J., Jedrzejowicz, P.: Gene expression programming as a data classification tool. A review. J. Intell. Fuzzy Syst. **36**(1), 91–100 (2019)
14. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997. pp. 179–186 (1997)
15. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. In: Quaglini, S., Barahona, P., Andreassen, S. (eds.) Artificial Intelligence Medicine, 8th Conference on AI in Medicine in Europe, AIME 2001, Cascais, Portugal, July 1-4, 2001, Proceedings. Lecture Notes in Computer Science, vol. 2101, pp. 63–66. Springer (2001)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62

16. Li, M., Xiong, A., Wang, L., Deng, S., Ye, J.: ACO resampling: Enhancing the performance of oversampling methods for class imbalance classification. Knowl. Based Syst. **196**, 105818 (2020)
17. Lin, W.C., Tsai, C.F., Hu, Y.H., Jhang, J.S.: Clustering-based undersampling in class-imbalanced data. Inf. Sci. **409**, 17–26 (2016)
18. López-García, P., Masegosa, A.D., Osaba, E., Onieva, E., Perallos, A.: Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics. Appl. Intell. **49**(8), 2807–2822 (2019)
19. Ofek, N., Rokach, L., Stern, R., Shabtai, A.: Fast-cbus: A fast clustering-based undersampling method for addressing the class imbalance problem. Neurocomputing **243**, 88–102 (2017)
20. Tang, S., ping Chen, S.: The generation mechanism of synthetic minority class examples. In: Proceedings of International Conference on Information Technology and Applications in Biomedicine. pp. 444–447 (2008)
21. Tomek, I.: Two modifications of cnn. IEEE Transactions on Systems, Man, and Cybernetics **SMC-6**(11), 769–772 (1976)
22. Tsai, C.F., Lin, W.C., Hu, Y.H., Yao, G.T.: Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf. Sci. **477**, 47–54 (2019)
23. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings. pp. 324–331 (2009)
24. Wang, Z., Li, Y., Li, D., Zhu, Z., Du, W.: Entropy and gravitation based dynamic radius nearest neighbor classification for imbalanced problem. Knowl. Based Syst. **193**, 105474 (2020)
25. Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-based learning algorithms. Mach. Learn. **38**(3), 257–286 (2000)
26. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Systems, Man, and Cybernetics **2**, 408–421 (1972)
27. Ye, X., Li, H., Imakura, A., Sakurai, T.: An oversampling framework for imbalanced classification based on laplacian eigenmaps. Neurocomputing **399**, 107–116 (2020)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_62