Fusers in homogenous ensemble of undersampled majority class for highly imbalanced data classification

Paweł Ksieniewicz PAWEL.KSIENIEWICZ@PWR.EDU.PL
Department of Systems and Computer Networks
Faculty of Electronics
Wrocław University of Science and Technology

Editor: Editor's name

Abstract

This is the abstract for this article.

Keywords: classification, classifier ensemble, undersampling, imbalanced data

1. Introduction

Additionally, in incremental learning, if the majority-class objects outnumber greatly the minority class, the latter can be completely ignored He and Garcia (2009). The aforementioned issues are reasons why most existing classification methods for imbalanced data are restricted to the *offline* learning only, i.e., a case where the entire data set is provided prior to the analysis.

Most of the classification algorithms assume that there are no significant disproportions among instances from different classes. Nevertheless, in many practical tasks, we may observe that instances from one class (so-called *majority class*) significantly outnumber the objects from remaining classes (*minority class*). Most of traditional classifiers have a bias in favor of the majority class although more often the minority class is more interesting, because misidentification of an instance belonging to it is usually much more expensive than assigning an instance from majority class to minority one. A good example is an undetected fraud that would be more expensive than the cost of additional analysis of a correct transaction classified as fraudless transaction. Such a problem is known as imbalanced data classification Sun et al. (2009); Wang et al. (2017), where an unequal number of instances from the examined classes plays a key role during the classifier learning. Various approaches have been proposed in the literature to tackle this challenging difficulty embedded in the nature of data. Usually, the researchers are focusing on maximizing the correct minority class classification. At the same time, performance on the majority class cannot be neglected.

In this project we will focus on binary imbalanced problems, because this setup is the one most frequently studied in the literature and most commonly meet in practical problems, e.g.,*fault detection* or *spam filtering*. Therefore, another important issue is proposing an appropriate quality measure that would be adequate for imbalanced data classification Elazmeh et al. (2006).

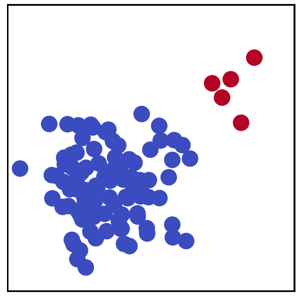


Figure 1: Easy separable imbalance dataset

In case of imbalanced data classification the disproportion between the different classes is not the sole issue of learning difficulties. One may easily came up with an example where the instance distributions from different classes are well-separated, as depicted in Fig.**??**.

Proposing a efficient classifier for such a task is not a challenge. Unfortunately, instances from the minority class often form clusters of an unknown structure that are scattered Napierala and Stefanowski (2012). Additional complication comes from the fact that during learning, the number of intactness from the minority class may be not sufficient enough for the learning algorithm to acquire the appropriate generalization level, which in effect can cause *overfitting* Chen and Wasikowski (2008). All those problems are a focus of intense research Chawla et al. (2002); Bunkhumpornpat et al. (2009); Kubat and Matwin (1997).

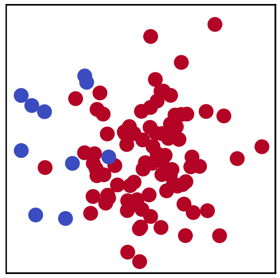
Methods for imbalanced data classification can be divided into three main groups Lopez et al. (2012).

Data preprocessing methods. This approach focuses on reducing the number of objects in majority class (*undersampling*) or generating new objects of the minority class (*oversampling*). The difference between *under-* and *oversampling* is presented in Fig. **??**.

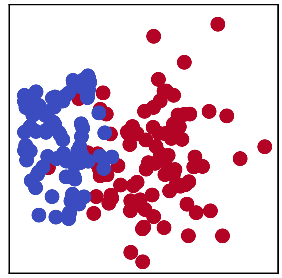
These mechanisms have the objective of balancing the quantity of instances from considered classes. For oversampling, new instances are random copies of existing ones or are generated in a guided manner. The most popular method is SMOTE Chawla et al. (2011) algorithm, which creates new instances on a basis of existing ones by slightly modifying the values of their attributes. As a result, new artificial examples that are in compliance with the minority class distribution are generated. Other oversampling methods are ADASYN He et al. (2008), in which a difficulty of an object for the classifying model is considered or RAMOBoost Chen et al. (2010). Unfortunately, methods such as SMOTE may lead to changes in the characteristic of the minority class and in result to *overfitting* the classifier, what was shown in Fig. **??**.

W WYPADKU UNDERSAMPLINGU NIE WYSTEPUJE RYZYKO NIESŁUSZNEGO ROZSZERZENIA PRZESTRZENI WZORCÓW KLASY MNIEJSZOŚCIOWEJ, CO MA MIEJSCE W CHOĆBY SMOTE.

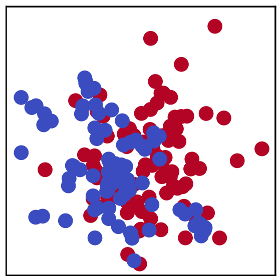
Several modifications of SMOTE have been proposed that are able to identify the instances to be copied in a more intelligent fashion such as *Borderline*SMOTE Han et al. (2005). It generates new instances from the minority class close to the decision border. *Safe-Level* SMOTE (Bunkhumpornpat et al., 2009) and LN-SMOTE Maciejewski and Stefanowski (2011) reduce the probability of generating synthetic instances of the minority class in areas where



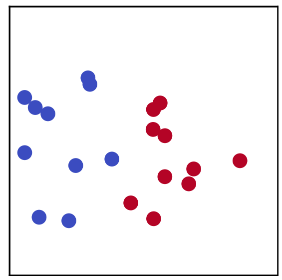
(*a*) Original dataset



(*b*) SMOTE

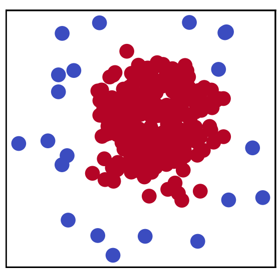


(*c*) ADASYN

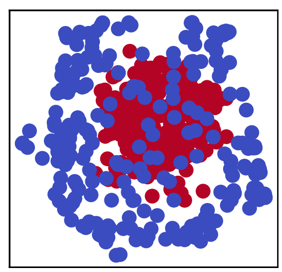


(*d*) Undersampling

Figure 2: Examples of data preprocessing methods.



(*a*) Original dataset



(*b*) SMOTE

Figure 3: Example of wrong SMOTE oversampling.

the predominant objects are that of the majority class. It is worth noticing that our team proposed two novel solutions to this problem: RBO Koziarski et al. (2017) and CCR that enforce instances from the majority-class to be relocated from the areas where the minority-class instances are present Koziarski and Woźniak (2017). Methods of *undersampling* are built around the idea of randomly removing the instances from the majority-class or removing them from the areas in such way that the quality of the classifier is not disrupted using neighbor analysis.

Inbuilt mechanisms. In this approach existing classification algorithms are adapted for imbalanced problems ensuring balanced accuracy for instances from both classes. Two of the most popular areas of research of this methods are using one-class classification Japkowicz

et al. (1995), usually known as learning without counterexamples, where the goal is to learn the minority class decision areas and because of the frequently assumed regular, closed shape of the decision borders is adequate to the clusters created by minority classes Krawczyk et al. (2014a). The disproportion between the number of instances in classes is then omitted. Another approach is the (*cost sensitive*) classification, where the algorithm takes into account the asymmetrical loss function that assigns a higher cost to a misclassification of an instance form a minority class Krawczyk et al. (2014b); Lopez et al. (2012); He and Garcia (2009); Zhou and Liu (2006). Unfortunately such methods can cause a reverse bias towards the minority class. Worth noting are methods based on ensemble classification Woźniak et al. (2014), like SMOTE*Boost* Chawla et al. (2003) and *AdaBoost.NC* Wang et al. (2010)
Hybrid methods. They combine the advantages of methods using data pre-processing with the classification methods. The most popular category is the hybridisation of *under-* and *oversampling* with ensemble classifiers Galar et al. (2012). This approach allows the data to be independently processed for each of the base model. Algorithms formed on modifications of *Bagging* and *Boosting* Chawla et al. (2003) enjoy wide popularity.

The main contributions of this work are:

2. Homogenous ensemble based on undersampling the majority class

Zaawansowane metody oversamplingu nie są możliwe do zastosowania przy sytuacji, gdzie w zbiorze uczącym znajduje się zaledwie kilka wzorców.

Idea k-foldowego podziału klasy większościowej. Wyznaczanie wartości k jako zaokrąglonego IR. Atut w postaci wykorzystania wszystkich wzorców, gdzie tworzymy komitet k zbalansowanych zbiorów.

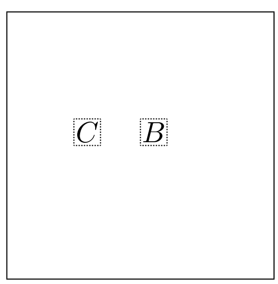


Figure 4: Scheme of using k-Fold division in ensemble construction

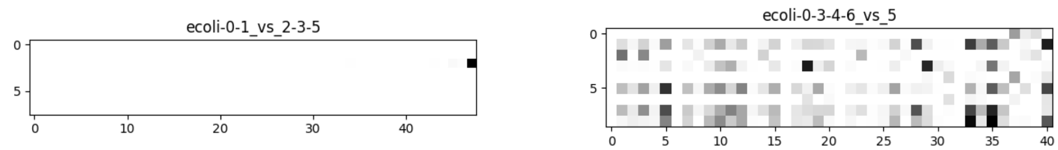
Wyliczanie wag. Accuracy się nie sprawdzi, więc BAC.

Jeśli klasyfikujemy nie jeden wzorzec, a wiele, wagi mogą być też dla pojedynczych próbek, dla podbicia, a więc pojawia się KONTRAST. Mamy takie ładne ilustracje z badań, dodajmy rysunek chociaż jeden poglądowo.

Potencjał kontrastu dla danych strumieniowych.

Proponowane metody decyzyjne.

- - akumulacja wsparć,
- - akumulacja ważona po członkach komitetu, gdzie waga to BAC dla zbioru uczącego,



(*a*) O jaki ładny (*b*) A ten brzydki

Figure 5: Rysunek.

- - akumulacja ważona po wzorcach, przez kontrast,
- - akumulacja znormalizowanych wag członków,
- - iloczyn znormalizowanych wag i kontrastu

Duża skala niezbalansowania to duża wielkość komitetu (ilustracja zależności na wykresie). Przyda się więc przycinanie (pruning).

Wyjaśnienie podejścia do pruningu. Wyliczamy wzajemną zależność statystyczną (Wilcoxonem) pomiędzy wsparciami członków i grupujemy – omijając kwestię 1z2 2z3 ale nie 1z3 – je uśredniając wsparcia w obrębie grupy. Uśredniamy też wagi i tworzymy tak dwupoziomowy system fuzji (potrzebna ilustracja).

Pruning też jest w kontekście klasyfikacji wielu wzorców na raz.

Wyjaśnienie kwestii wspomnianej wcześniej i uzasadnienie pominięcia jej analizy.

3. Experiment design

Wybrane zbiory danych.

Wykorzystane klasyfikatory bazowe. Wyjaśnienie dlaczego odrzuciliśmy MLP (brak konwergencji na bardzo niewielkich zbiorach) i SVC (nie jest on naturalnie probabilistyczny, a jego probabilistyczna interpretacja jest silnie zakłamana przy niewielkich zbiorach danych). Stąd bierzemy GNB, kNN i DT, przy domyślnych parametrach z sklearn.

Powównawczo uczenie na pełnym zbiorze i zbiorach po pojedynczym under i oversamplingu.

Undersampling, ze względu na niestabilność, powtórzony pięciokrotnie na każdym foldzie.

Zastosowana metoda podziału – wymuszone przez KEEL k-fold CV (z k=5).

Zastosowana miara jakości – zbalansowana dokładność, wymierna w niezbalansowanych danych.

Zastosowana analiza statystyczna – parowa zależność pomiędzy klasyfikatorem z najwyższym rezultatem a pozostałymi w postaci testu Wilcoxona.

Przygotowane oprogramowanie ze wskazaniem repozytorium.

4. Experimental evaluation

Przedstawienie tabel.

Tabela zbiorcza zwycięstw w zależności od parametrów (z grupowaniem).

Without oversampled set										With oversampled set										OS	US	Full	Dataset
All members					Reduced members					All members					Reduced members								
REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC				
.845	.843	.843	.837	.837	.845	.845	.845	.837	.837	.841	.841	.839	.828	.825	.845	.845	.845	.837	.837	.806	.838	.825	*ecoli-0-1-3-7-vs-2-6*
.809	.834	.859	.879	.857	.748	.800	.820	.926	.898	.874	.898	.868	.913	.904	.746	.896	.895	.923	.922	.859	.787	.878	*ecoli4*
.580	.580	.610	.580	.610	.577	.577	.577	.583	.580	.552	.552	.552	.560	.591	.577	.577	.577	.585	.585	.569	.589	.580	*glass-0-1-6-vs-2*
.989	.989	.989	.939	.941	.989	.989	.989	.989	.989	.941	.941	.941	.941	.941	.989	.989	.989	.989	.989	.941	.975	.941	*glass-0-1-6-vs-5*
.619	.619	.619	.616	.616	.641	.641	.641	.641	.641	.619	.616	.619	.616	.616	.641	.641	.644	.641	.641	.617	.620	.591	*glass2*
.800	.797	.800	.802	.802	.807	.807	.804	.766	.766	.781	.781	.774	.779	.779	.812	.774	.774	.771	.771	.731	.745	.587	*glass4*
.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.945	.938	*glass5*
.817	.817	.817	.846	.845	.828	.828	.828	.828	.828	.807	.807	.803	.846	.845	.831	.831	.831	.830	.828	.791	.806	.763	*page-blocks-1-3-vs-4*
.994	.994	.994	.995	.994	.994	.994	.994	.991	.991	.995	.995	.995	.991	.991	.994	.994	.994	.991	.991	.990	.993	.991	*shuttle-c0-vs-c4*
.971	.975	.975	.979	.979	.996	.996	.996	.996	.996	.984	.979	.979	.988	.988	.996	.996	.996	.996	.996	.988	.946	.996	*shuttle-c2-vs-c4*
.909	.909	.909	.902	.903	.909	.909	.909	.909	.909	.910	.910	.910	.904	.903	.910	.909	.909	.909	.909	.906	.905	.917	*vowel0*
.690	.706	.689	.728	.731	.680	.737	.727	.786	.774	.702	.700	.677	.726	.715	.687	.724	.710	.785	.791	.498	.601	.504	*yeast-0-5-6-7-9-vs-4*
.563	.570	.567	.576	.570	.562	.557	.562	.627	.642	.563	.552	.553	.574	.566	.556	.564	.556	.625	.639	.540	.598	.544	*yeast-1-2-8-9-vs-7*
.552	.553	.550	.557	.563	.567	.566	.554	.550	.575	.551	.558	.554	.555	.590	.588	.555	.562	.564	.568	.541	.566	.547	*yeast-1-4-5-8-vs-7*
.698	.698	.699	.705	.703	.703	.710	.689	.726	.722	.671	.681	.674	.705	.700	.700	.695	.674	.725	.719	.586	.686	.604	*yeast-1-vs-7*
.804	.801	.803	.799	.800	.820	.822	.827	.870	.862	.773	.794	.781	.799	.800	.805	.833	.827	.869	.861	.529	.739	.561	*yeast-2-vs-4*
.773	.773	.773	.773	.774	.773	.773	.773	.773	.773	.796	.771	.796	.773	.773	.773	.773	.773	.773	.773	.616	.762	.657	*yeast-2-vs-8*
.675	.674	.674	.769	.728	.708	.766	.766	.822	.813	.644	.655	.637	.781	.746	.710	.765	.769	.820	.811	.526	.660	.551	*yeast4*
.934	.934	.934	.934	.934	.929	.934	.938	.957	.955	.917	.921	.919	.934	.934	.927	.930	.935	.957	.955	.780	.910	.831	*yeast5*
.785	.787	.787	.847	.796	.829	.863	.848	.887	.878	.760	.773	.768	.843	.795	.818	.860	.845	.887	.878	.628	.795	.650	*yeast6*
.785	.781	.783	.804	.804	.660	.810	.810	.906	.913	.810	.812	.860	.906	.906	.785	.837	.862	.910	.913	.885	.679	.877	*ecoli-0-1-4-6-vs-5*
.630	.630	.630	.630	.630	.630	.630	.630	.652	.654	.630	.630	.630	.662	.663	.630	.630	.630	.657	.659	.667	.634	.630	*ecoli-0-1-4-7-vs-2-3-5-6*
.617	.617	.617	.710	.693	.617	.617	.617	.823	.829	.637	.677	.755	.852	.855	.617	.617	.617	.826	.831	.863	.668	.735	*ecoli-0-1-4-7-vs-5-6*
.578	.578	.578	.575	.578	.558	.578	.578	.658	.658	.618	.618	.638	.638	.638	.558	.618	.618	.658	.658	.639	.578	.638	*ecoli-0-1-vs-2-3-5*
.639	.689	.689	.736	.732	.616	.666	.664	.741	.741	.714	.739	.789	.834	.830	.641	.691	.689	.816	.816	.797	.658	.782	*ecoli-0-1-vs-5*
.683	.733	.708	.778	.753	.658	.733	.733	.853	.853	.756	.758	.806	.828	.803	.731	.758	.758	.895	.895	.638	.657	.754	*ecoli-0-2-3-4-vs-5*
.563	.563	.588	.635	.613	.588	.588	.588	.625	.630	.588	.588	.588	.625	.628	.588	.613	.613	.623	.628	.592	.595	.563	*ecoli-0-2-6-7-vs-3-5*
.876	.876	.876	.876	.876	.818	.870	.873	.901	.898	.859	.865	.849	.857	.851	.843	.870	.870	.901	.895	.725	.716	.784	*ecoli-0-3-4-6-vs-5*
.633	.633	.633	.765	.729	.633	.673	.653	.781	.799	.697	.735	.726	.785	.747	.653	.673	.673	.789	.791	.734	.665	.775	*ecoli-0-3-4-7-vs-5-6*
.756	.756	.756	.833	.792	.711	.786	.783	.883	.883	.769	.794	.811	.772	.758	.758	.783	.783	.883	.883	.730	.657	.817	*ecoli-0-3-4-vs-5*
.876	.876	.876	.901	.903	.853	.876	.878	.903	.878	.901	.901	.901	.901	.901	.878	.878	.876	.903	.901	.890	.725	.854	*ecoli-0-4-6-vs-5*
.598	.598	.598	.573	.555	.548	.548	.548	.640	.597	.557	.557	.557	.575	.575	.548	.508	.508	.618	.597	.544	.571	.508	*ecoli-0-6-7-vs-3-5*
.795	.787	.793	.805	.812	.685	.755	.705	.853	.830	.845	.838	.838	.828	.825	.688	.755	.755	.853	.830	.847	.682	.780	*ecoli-0-6-7-vs-5*
.558	.589	.558	.615	.581	.622	.622	.620	.620	.620	.592	.592	.558	.595	.595	.622	.622	.620	.620	.620	.597	.590	.577	*glass-0-1-4-6-vs-2*
.567	.582	.582	.527	.534	.542	.536	.542	.604	.582	.527	.527	.527	.530	.533	.533	.551	.558	.584	.584	.508	.555	.519	*glass-0-1-5-vs-2*
.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.984	.984	.994	*glass-0-4-vs-5*
.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.995	.945	.989	.945	*glass-0-6-vs-5*
.576	.576	.576	.576	.576	.576	.576	.576	.782	.782	.634	.657	.692	.777	.768	.576	.576	.576	.781	.783	.782	.605	.670	*yeast-0-2-5-6-vs-3-7-8-9*
.883	.893	.889	.901	.901	.901	.901	.902	.897	.900	.779	.834	.841	.900	.898	.894	.896	.896	.897	.900	.524	.785	.577	*yeast-0-2-5-7-9-vs-3-6-8*
.597	.632	.606	.620	.621	.606	.605	.606	.605	.605	.601	.607	.589	.619	.633	.611	.612	.600	.605	.605	.539	.633	.557	*yeast-0-3-5-9-vs-7-8*

Table 1: Balanced accuracy scores obtained using GNB as a base classifier

Dataset	Without oversampled set										With oversampled set										OS	US	Full
	All members					Reduced members					All members					Reduced members							
	REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC			
ecoli-0-1-3-7-vs-2-6	.836	.836	.836	.847	.849	.845	.845	.847	.844	.845	.830	.830	.832	.835	.834	.854	.853	.851	.853	.856	.835	.835	.850
ecoli4	.941	.941	.945	.949	.949	.940	.940	.940	.921	.945	.964	.967	.972	.909	.909	.945	.945	.945	.940	.943	.909	.928	.848
glass-0-1-6-vs-2	.718	.718	.718	.715	.724	.732	.735	.726	.746	.699	.717	.751	.760	.750	.728	.705	.705	.737	.758	.735	.656	.666	.555
glass-0-1-6-vs-5	.880	.880	.880	.883	.883	.830	.877	.880	.880	.880	.836	.839	.853	.879	.879	.833	.833	.836	.833	.841	.933	.852	.739
glass2	.696	.696	.724	.721	.719	.701	.724	.718	.724	.724	.737	.757	.746	.756	.756	.726	.732	.731	.723	.746	.715	.678	.485
glass4	.861	.861	.861	.878	.875	.861	.861	.861	.863	.863	.868	.868	.888	.905	.913	.861	.861	.861	.888	.898	.925	.865	.781
glass5	.823	.823	.826	.830	.830	.813	.813	.813	.813	.816	.838	.843	.879	.801	.873	.816	.813	.826	.821	.833	.830	.811	.695
page-blocks-1-3-vs-4	.868	.868	.874	.877	.877	.867	.871	.866	.902	.897	.901	.907	.926	.909	.924	.874	.875	.881	.917	.920	.917	.872	.808
shuttle-c0-vs-c4	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996
shuttle-c2-vs-c4	.859	.859	.859	.891	.859	.879	.883	.850	.883	.850	.996	.000	.000	.000	.000	.887	.887	.906	.927	.996	.000	.845	.600
vowel0	.942	.942	.942	.944	.943	.943	.942	.943	.945	.945	.952	.953	.963	.997	.996	.948	.948	.954	.980	.986	.999	.939	.977
yeast-0-5-6-7-9-vs-4	.781	.781	.785	.786	.790	.779	.779	.785	.785	.787	.810	.824	.832	.830	.820	.799	.800	.819	.828	.829	.795	.792	.667
yeast-1-2-8-9-vs-7	.644	.645	.662	.665	.663	.659	.660	.655	.645	.659	.702	.695	.673	.625	.626	.655	.660	.684	.692	.693	.627	.652	.499
yeast-1-4-5-8-vs-7	.576	.574	.597	.593	.592	.571	.571	.602	.571	.584	.622	.628	.614	.612	.609	.594	.607	.623	.636	.643	.615	.590	.499
yeast-1-vs-7	.698	.699	.695	.705	.697	.730	.729	.736	.733	.742	.737	.743	.712	.700	.700	.745	.755	.734	.706	.710	.705	.682	.517
yeast-2-vs-4	.915	.915	.914	.907	.908	.914	.914	.912	.917	.917	.921	.921	.913	.906	.906	.921	.920	.921	.921	.909	.885	.908	.819
yeast-2-vs-8	.715	.715	.740	.719	.719	.721	.721	.728	.727	.734	.782	.782	.778	.803	.803	.738	.738	.752	.755	.758	.803	.734	.774
yeast4	.819	.819	.819	.839	.838	.840	.841	.841	.841	.843	.838	.829	.839	.771	.774	.844	.844	.843	.841	.852	.749	.835	.574
yeast5	.955	.955	.954	.955	.954	.957	.957	.956	.957	.958	.956	.956	.960	.962	.967	.958	.958	.958	.960	.960	.964	.952	.850
yeast6	.883	.883	.884	.886	.885	.886	.886	.885	.887	.886	.899	.900	.892	.840	.865	.887	.887	.888	.893	.895	.840	.879	.739
ecoli-0-1-4-6-vs-5	.890	.890	.890	.892	.892	.892	.892	.892	.892	.890	.894	.894	.896	.904	.906	.892	.892	.892	.892	.892	.917	.886	.898
ecoli-0-1-4-7-vs-2-3-5-6	.888	.888	.888	.890	.885	.893	.895	.891	.896	.896	.869	.869	.886	.869	.886	.886	.886	.889	.904	.902	.856	.882	.847
ecoli-0-1-4-7-vs-5-6	.886	.886	.888	.868	.868	.883	.885	.881	.913	.891	.874	.876	.887	.892	.894	.893	.893	.871	.887	.889	.899	.883	.838
ecoli-0-1-vs-2-3-5	.892	.892	.892	.914	.894	.892	.892	.872	.892	.872	.872	.865	.867	.885	.885	.874	.874	.872	.883	.863	.887	.895	.830
ecoli-0-1-vs-5	.905	.905	.907	.914	.911	.905	.905	.907	.911	.911	.911	.911	.911	.911	.914	.909	.909	.911	.911	.911	.916	.902	.900
ecoli-0-2-3-4-vs-5	.900	.900	.900	.895	.897	.900	.900	.903	.900	.903	.900	.900	.900	.900	.900	.900	.900	.900	.900	.900	.909	.904	.894
ecoli-0-2-6-7-vs-3-5	.836	.836	.836	.841	.841	.833	.836	.833	.816	.811	.853	.858	.860	.830	.833	.843	.843	.848	.860	.855	.890	.814	.787
ecoli-0-3-4-6-vs-5	.887	.887	.874	.890	.890	.887	.887	.879	.890	.890	.893	.893	.895	.901	.901	.893	.893	.890	.895	.895	.911	.881	.875
ecoli-0-3-4-7-vs-5-6	.904	.904	.902	.917	.917	.904	.904	.902	.921	.915	.901	.901	.901	.890	.890	.913	.911	.897	.885	.888	.894	.887	.876
ecoli-0-3-4-vs-5	.903	.903	.903	.903	.900	.900	.900	.903	.903	.900	.906	.906	.906	.903	.903	.906	.906	.906	.903	.903	.911	.889	.875
ecoli-0-4-6-vs-5	.890	.890	.890	.887	.884	.884	.884	.884	.890	.890	.890	.890	.890	.898	.895	.890	.890	.890	.892	.895	.914	.889	.900
ecoli-0-6-7-vs-3-5	.860	.860	.860	.872	.875	.857	.857	.855	.872	.867	.875	.875	.855	.865	.863	.865	.870	.870	.863	.858	.893	.844	.835
ecoli-0-6-7-vs-5	.853	.853	.850	.868	.863	.860	.860	.858	.870	.872	.878	.878	.880	.863	.885	.865	.865	.872	.890	.887	.863	.850	.847
glass-0-1-4-6-vs-2	.715	.715	.706	.712	.725	.717	.712	.712	.714	.717	.731	.717	.701	.718	.716	.741	.749	.747	.734	.727	.732	.681	.512
glass-0-1-5-vs-2	.684	.650	.653	.663	.653	.674	.674	.650	.653	.650	.705	.728	.748	.767	.740	.676	.695	.712	.745	.758	.656	.651	.527
glass-0-4-vs-5	.932	.932	.939	.932	.932	.939	.939	.944	.939	.944	.951	.951	.951	.951	.951	.951	.951	.951	.951	.951	.988	.917	.850
glass-0-6-vs-5	.873	.873	.873	.868	.868	.868	.868	.873	.878	.878	.883	.898	.928	.953	.953	.878	.878	.878	.913	.933	.985	.816	.745
yeast-0-2-5-6-vs-3-7-8-9	.784	.784	.786	.781	.784	.785	.785	.790	.786	.790	.801	.802	.800	.791	.798	.793	.798	.800	.795	.798	.784	.760	.762
yeast-0-2-5-7-9-vs-3-6-8	.909	.909	.908	.911	.911	.904	.904	.908	.906	.910	.913	.913	.902	.900	.900	.909	.913	.912	.896	.895	.904	.902	.902
yeast-0-3-5-9-vs-7-8	.757	.757	.755	.746	.753	.748	.748	.757	.749	.754	.759	.742	.739	.731	.730	.770	.767	.750	.723	.734	.718	.702	.639

Table 2: Balanced accuracy scores obtained using kNN as a base classifier

Without oversampled set										With oversampled set													
All members					Reduced members					All members					Reduced members					OS	US	Full	Dataset
REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC	REG	WEI	CON	NOR	NC				
.816	.816	.816	.818	.816	.823	.823	.823	.838	.838	.783	.784	.784	.715	.725	.823	.823	.823	.842	.842	.624	.708	.841	*ecoli-0-1-3-7-vs-2-6*
.866	.866	.866	.866	.866	.876	.873	.873	.873	.873	.876	.878	.878	.865	.889	.876	.876	.876	.834	.834	.817	.848	.866	*ecoli4*
.740	.737	.731	.657	.657	.732	.727	.727	.680	.680	.716	.716	.722	.603	.584	.741	.738	.738	.700	.700	.581	.630	.546	*glass-0-1-6-vs-2*
.929	.929	.929	.937	.934	.934	.934	.934	.937	.937	.934	.937	.937	.943	.943	.934	.934	.934	.940	.940	.859	.886	.936	*glass-0-1-6-vs-5*
.739	.742	.739	.803	.757	.785	.780	.780	.784	.784	.749	.684	.684	.634	.644	.767	.800	.800	.801	.801	.616	.682	.573	*glass2*
.885	.885	.885	.898	.893	.900	.897	.897	.907	.907	.898	.898	.895	.868	.868	.905	.903	.903	.860	.860	.819	.835	.804	*glass4*
.924	.924	.924	.879	.877	.934	.934	.934	.891	.891	.939	.939	.946	.971	.973	.939	.939	.939	.949	.949	.933	.867	.898	*glass5*
.982	.982	.981	.987	.987	.991	.991	.991	.992	.992	.989	.989	.990	.993	.993	.991	.991	.991	.992	.992	.994	.958	.996	*page-blocks-1-3-vs-4*
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	*shuttle-c0-vs-c4*
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.950	.950	.950	.950	.950	.000	.000	.000	.000	.000	.990	.959	.950	*shuttle-c2-vs-c4*
.955	.955	.955	.956	.956	.957	.956	.956	.956	.956	.959	.959	.958	.951	.951	.957	.957	.957	.961	.961	.921	.940	.936	*vowel0*
.788	.788	.788	.783	.786	.787	.787	.787	.787	.787	.760	.758	.751	.704	.705	.778	.777	.777	.759	.759	.674	.750	.659	*yeast-0-5-6-7-9-vs-4*
.731	.733	.729	.722	.737	.736	.736	.736	.732	.732	.733	.742	.738	.659	.647	.749	.748	.748	.750	.750	.621	.624	.630	*yeast-1-2-8-9-vs-7*
.637	.639	.643	.621	.602	.633	.624	.624	.586	.586	.656	.650	.645	.570	.562	.649	.649	.649	.587	.587	.533	.581	.537	*yeast-1-4-5-8-vs-7*
.767	.768	.764	.784	.781	.783	.794	.794	.801	.801	.737	.741	.709	.622	.598	.782	.782	.782	.789	.789	.603	.661	.683	*yeast-1-vs-7*
.950	.950	.951	.955	.955	.948	.948	.948	.954	.954	.958	.958	.959	.895	.888	.960	.958	.958	.931	.931	.822	.900	.843	*yeast-2-vs-4*
.798	.795	.794	.778	.774	.787	.787	.787	.782	.782	.808	.813	.820	.737	.716	.790	.789	.789	.802	.802	.697	.715	.690	*yeast-2-vs-8*
.824	.824	.817	.837	.826	.846	.846	.846	.844	.844	.815	.818	.830	.651	.642	.850	.850	.850	.845	.845	.626	.792	.643	*yeast4*
.962	.962	.961	.965	.964	.964	.964	.964	.965	.965	.966	.967	.956	.931	.932	.964	.964	.964	.967	.967	.845	.936	.845	*yeast5*
.838	.838	.839	.840	.838	.847	.847	.847	.852	.852	.850	.851	.856	.804	.782	.852	.851	.851	.860	.860	.750	.818	.730	*yeast6*
.900	.900	.898	.873	.873	.910	.910	.910	.887	.887	.885	.885	.885	.871	.871	.892	.887	.887	.860	.860	.794	.823	.781	*ecoli-0-1-4-6-vs-5*
.842	.842	.838	.819	.819	.847	.845	.845	.830	.830	.848	.849	.849	.869	.852	.848	.848	.848	.866	.866	.827	.804	.820	*ecoli-0-1-4-7-vs-2-3-5-6*
.866	.866	.866	.853	.853	.878	.871	.871	.876	.876	.879	.881	.881	.867	.867	.876	.876	.876	.867	.867	.844	.803	.787	*ecoli-0-1-4-7-vs-5-6*
.873	.871	.871	.873	.873	.878	.878	.878	.865	.865	.840	.840	.840	.838	.838	.820	.820	.820	.831	.831	.764	.802	.760	*ecoli-0-1-vs-2-3-5*
.873	.873	.873	.873	.873	.848	.848	.848	.850	.850	.850	.850	.850	.855	.830	.852	.850	.850	.855	.855	.805	.841	.857	*ecoli-0-1-vs-5*
.914	.911	.911	.909	.909	.911	.911	.911	.911	.911	.917	.917	.917	.897	.895	.917	.917	.917	.897	.897	.832	.843	.781	*ecoli-0-2-3-4-vs-5*
.823	.823	.823	.820	.820	.833	.823	.823	.820	.820	.833	.833	.833	.838	.838	.833	.830	.830	.835	.835	.811	.791	.790	*ecoli-0-2-6-7-vs-3-5*
.882	.882	.884	.901	.904	.879	.879	.879	.887	.887	.893	.893	.893	.893	.893	.906	.901	.901	.901	.901	.812	.834	.786	*ecoli-0-3-4-6-vs-5*
.875	.875	.875	.853	.853	.875	.875	.875	.853	.853	.866	.866	.866	.859	.859	.868	.864	.864	.851	.851	.836	.839	.840	*ecoli-0-3-4-7-vs-5-6*
.928	.928	.928	.914	.914	.928	.928	.928	.911	.911	.936	.936	.936	.944	.944	.911	.936	.936	.947	.947	.869	.862	.831	*ecoli-0-3-4-vs-5*
.912	.912	.912	.912	.912	.906	.906	.906	.912	.912	.906	.906	.906	.829	.831	.906	.904	.904	.859	.859	.813	.838	.836	*ecoli-0-4-6-vs-5*
.818	.818	.818	.825	.825	.818	.818	.818	.830	.830	.832	.832	.840	.865	.867	.840	.835	.835	.853	.853	.864	.786	.850	*ecoli-0-6-7-vs-3-5*
.867	.867	.867	.875	.875	.882	.875	.875	.878	.878	.863	.863	.863	.838	.812	.880	.880	.880	.838	.838	.827	.819	.795	*ecoli-0-6-7-vs-5*
.835	.833	.830	.805	.810	.835	.827	.827	.746	.746	.847	.813	.796	.711	.681	.739	.739	.739	.715	.715	.676	.675	.610	*glass-0-1-4-6-vs-2*
.806	.803	.803	.836	.840	.734	.800	.800	.819	.819	.704	.704	.704	.557	.557	.754	.751	.751	.535	.535	.572	.634	.578	*glass-0-1-5-vs-2*
.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.982	.994	.942	.994	*glass-0-4-vs-5*
.959	.959	.959	.974	.974	.959	.959	.959	.980	.980	.964	.964	.969	.985	.985	.975	.975	.975	.990	.990	.955	.879	.995	*glass-0-6-vs-5*
.775	.775	.770	.758	.757	.782	.782	.782	.790	.790	.800	.802	.805	.742	.723	.801	.795	.795	.780	.780	.701	.732	.733	*yeast-0-2-5-6-vs-3-7-8-9*
.903	.903	.903	.888	.887	.898	.898	.898	.895	.895	.896	.896	.896	.876	.878	.903	.900	.900	.897	.897	.867	.868	.854	*yeast-0-2-5-7-9-vs-3-6-8*
.736	.736	.737	.709	.716	.729	.729	.729	.724	.724	.728	.724	.724	.627	.628	.720	.715	.715	.651	.651	.599	.635	.688	*yeast-0-3-5-9-vs-7-8*

Table 3: Balanced accuracy scores obtained using DTC as a base classifier

Interpretacja wyników, czyli co zostało należycie uprawdopodobnione.

5. Conclusions

Co zostało zaproponowane.

Na co pozwala taka metoda.

Do jakich rezultatów doprowadziła.

Jakie są plany na przyszłość (czyli co robisz w wakacje).

Acknowledgments

Acknowledgements go here.

References

C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Safe-Level-SMOTE: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference 2009, Bangkok, Thailand, April 27-30, 2009, Proceedings*, pages 475–482, 2009.

N V Chawla, K W Bowyer, L O Hall, and W P Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *arXiv.org*, June 2011.

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.

Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. *SMOTEBoost: Improving Prediction of the Minority Class in Boosting*, pages 107–119. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-39804-2. doi: 10.1007/978-3-540-39804-2_12. URL https://doi.org/10.1007/978-3-540-39804-2_12.

S. Chen, H. He, and E. A. Garcia. RAMOBoost: Ranked minority oversampling in boosting. *IEEE Transactions on Neural Networks*, 21(10):1624–1642, 2010.

Xue-wen Chen and Michael Wasikowski. Fast: A ROC-based feature selection metric for small samples and imbalanced data classification problems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 124–132, 2008.

William Elazmeh, Nathalie Japkowicz, and Stan Matwin. Evaluating misclassifications in imbalanced data. In *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, pages 126–137, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-45375-X, 978-3-540-45375-8. doi: 10.1007/11871842_16. URL http://dx.doi.org/10.1007/11871842_16.

M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. 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, July 2012. ISSN 1094-6977. doi: 10.1109/TSMCC.2011.2161285.

H. Han, W. Wang, and B. Mao. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing, International Conference on Intelligent Computing 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I*, pages 878–887, 2005.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the International Joint Conference on Neural Networks, 2008, part of the IEEE World Congress on Computational Intelligence, 2008, Hong Kong, China, June 1-6, 2008*, pages 1322–1328, 2008.

Nathalie Japkowicz, Catherine Myers, and Mark Gluck. A novelty detection approach to classification. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 518–523, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8, 978-1-558-60363-9. URL http://dl.acm.org/citation.cfm?id=1625855.1625923.

Michał Koziarski, Bartosz Krawczyk, and Michał Woźniak. Radial-based approach to imbalanced data oversampling. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 318–327. Springer, 2017.

Michał Koziarski and Michał Woźniak. Ccr: Combined cleaning and resampling algorithm for imbalanced data classification. *International Journal of Applied Mathematics and Computer Science*, 27(4), 2017.

Bartosz Krawczyk, Michal Wozniak, and Boguslaw Cyganek. Clustering-based ensembles for one-class classification. *Information Sciences*, 264:182–195, 2014a.

Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14(Part C):554 – 562, 2014b. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2013.08.014. URL http://www.sciencedirect.com/science/article/pii/S1568494613002895.

Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the 14th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.

V. Lopez, A. Fernandez, J. G. Moreno-Torres, and F. Herrera. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608, 2012.

T. Maciejewski and J. Stefanowski. Local neighbourhood extension of SMOTE for mining imbalanced data. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining 2011, part of the IEEE Symposium Series on Computational Intelligence 2011, April 11-15, 2011, Paris, France*, pages 104–111, 2011.

K. Napierala and J. Stefanowski. Identification of different types of minority class examples in imbalanced data. In *Hybrid Artificial Intelligent Systems*, volume 7209 of *Lecture Notes in Computer Science*, pages 139–150. Springer Berlin Heidelberg, 2012.

Y. Sun, A. K. C. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009.

S. Wang, H. Chen, and X. Yao. Negative correlation learning for classification ensembles. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2010. doi: 10.1109/IJCNN.2010.5596702.

Shuo Wang, Leandro L. Minku, and Xin Yao. A systematic study of online class imbalance learning with concept drift. *CoRR*, abs/1703.06683, 2017. URL http://arxiv.org/abs/1703.06683.

Michał Woźniak, Manuel Graña, and Emilio Corchado. A survey of multiple classifier systems as hybrid systems. *Inf. Fusion*, 16:3–17, March 2014. ISSN 1566-2535. doi: 10.1016/j.inffus.2013.04.006. URL http://dx.doi.org/10.1016/j.inffus.2013.04.006.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006. ISSN 1041-4347. doi: 10.1109/TKDE.2006.17.