Proceedings of Machine Learning Research 1:1–10, 2010 Workshop Title

Homogenous ensemble of undersampled majority class for highly imbalanced data binary 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

The main contributions of this work are:

- metoda konstrukcji komitetu na k-fold,
- propozycje reguł decyzyjnych,
- metoda pruningu dostosowującego regułę decyzyjną do zbioru testowego
- implementacja i ewaluacja eksperymentalna

2. Homogenous ensemble based on undersampling the majority class

2.1. Establishing ensemble

Complex oversampling methods, such as SMOTE or ADASYN, despite the large possibilities in most of the imbalanced problems, are not applicable to extreme situations where the minority class is represented by only a few samples, which makes it impossible to designate the nearest neighbors to create a new synthetic object. This could lead to the use of *undersampling* in such problems, but it is characterized, due to high randomness, by a strong instability in a situation of high IR (*imbalance ratio*), which does not allow for the development of a reliable solution.

A popular answer to the above-mentioned problem are the ensemble methods of *Bagging* or *Boosting*, characterized by random sampling with replacement of the training set, breaking a large problem, into a set of smaller problems. This work proposes a basic method, which also breaks the imbalanced task, but with ensuring the use of all the patterns available in the data set, but without a risk of overlapping. Its description can be found in Algorithm 1.

Algorithm 1: Training classifier ensemble from multiple balanced training datasets separated from one imbalanced dataset of binary problem

Given a dataset DS:

1. Divide DS into subsets of minority- $MinC$ and majority-class $MajC$
2. Calculate imbalanced ratio IR as the proportion of the number of patterns in $MinC$ and $MajC$
3. Establish k by rounding IR to nearest integer
4. Perform a *shuffled k-fold division* of $MajC$ to produce a set of subsets $MajC_1, MajC_2, \ldots, MajC_k$
5. For every i in range to k
    6. Join $MajC_i$ with $MinC$ to prepare a training set TS_i,
    7. Train classifier Ψ_i on TS_i and add it into ensemble

After dividing the dataset with imbalanced binary problem into separated minority ($MinC$) and majority class ($MajC$), we are calculating the IR (*imbalanced ratio*) between given classes. Rounding IR to the nearest integer value k allows us to find the optimal division coefficient of the majority class samples in the context of maximizing the balance between the $MinC$ and any $MajC_i$ subsets while ensuring that all $MajC$ patterns are used in learning process with no overlapping between the individual $MajC_i$'s. Each of k classifiers Ψ_i is trained on union of $MinC$ and $MajC_i$ sets.

Extending pool with oversampling As an extension of the method of classifier ensemble construction, it is also proposed to extend its pool by a model learned on an additional data set, which is a full set of data subjected to *oversampling*. It is worth testing if the knowledge gained from this method may be a valuable contribution to the ensemble decision. Due to impossibility to use SMOTE or ADASYN for oversampling the minority class with only few instances, only its basic variant will be used.

2.2. Fuser design

In addition to ensuring the diversity of the classifiers pool, which we achieve by a homogenous committee built on disjoint subsets of the majority class supplemented by minority patterns, the key aspect of the hybrid classification system is the appropriate design of its *fuser* – the element responsible for making decisions based on the answers of the base classifiers.

There are two groups of solutions here. The first are based on component *decisions* of the committee, most often employing the *majority voting* to produce a final decision. The decision rules proposed in this work are, however, part of the second group, where the *fuser* is carried out by *averaging* (or *accumulating*) the *support vectors* received from the members of a pool.

Note:
It should be remembered that in such methods, it is necessary to use a *probabilistic classification model*, which also requires *quantitative* and not *qualitative data*.

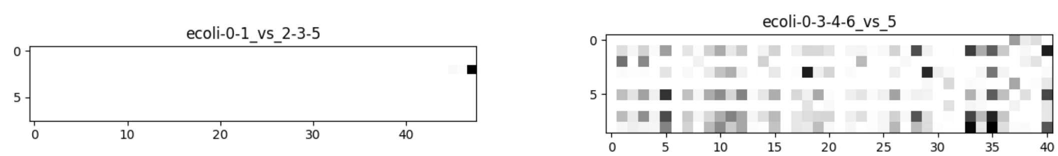
Five fusers were proposed:

1. **reg** — regular accumulation of support, without weighing the members of the committee.
2. **wei** — accumulation weighted after members of the committee.

   The weight of the classifier in the pool is its quality achieved for the training set. We can not use here the measure of *accuracy*, which does not fit with the task of the imbalanced classification, so we decided on a *balanced accuracy* (Brodersen et al., 2010).
3. **nor** — akumulacja znormalizowanych wag członków,
4. - con akumulacja ważona po wzorcach, przez kontrast,
5. - nci iloczyn znormalizowanych wag i kontrastu

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.



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

Figure 1: Rysunek.

Potencjał kontrastu dla danych strumieniowych.

Proponowane metody decyzyjne.

W konstrukcji reguły decyzyjnej opieramy się na wsparciu dla klasy pozytywnej.

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

2.3. Ensemble 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.

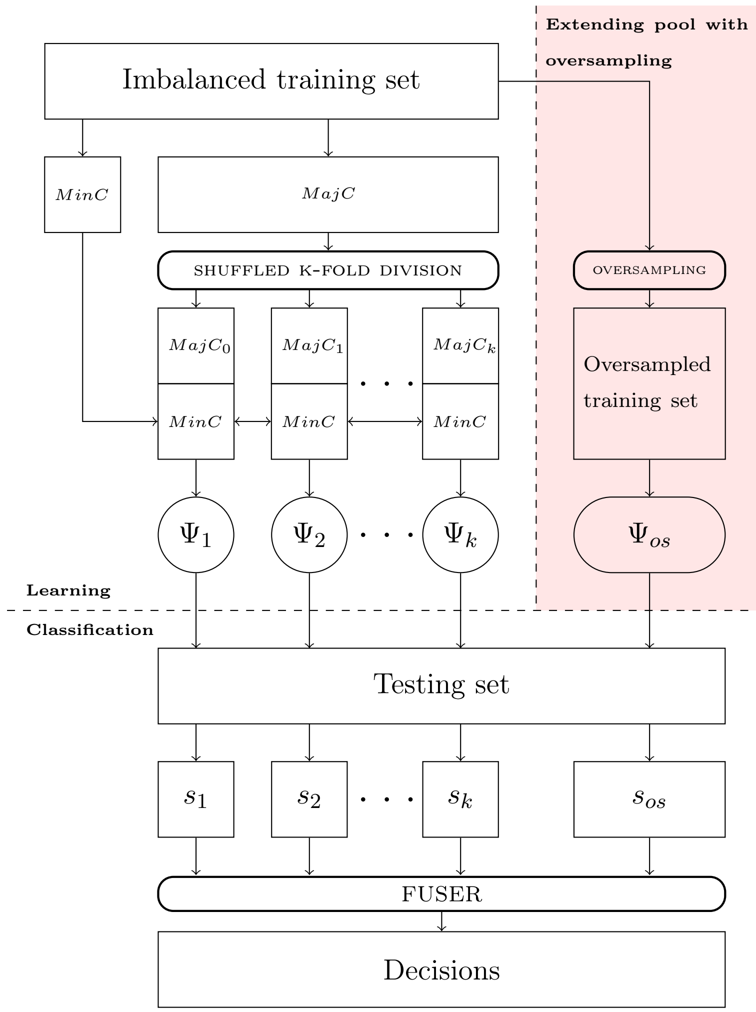


Figure 2: Scheme of using k-Fold division in ensemble construction

3. Experiment design

For the experimental evaluation of the proposed method, a collection of datasets made available with KEEL (Alcalá-Fdez et al., 2011) was used, focusing on a section containing highly unbalanced data, with IR greater than 9 (Fernández et al., 2009). From among the available datasets, 40 were selected presenting only binary problems with quantitative attributes. A review of selected datasets, including information on their number of features, the number of patterns in each class and the unbalance ratio is presented in Table 1.

As may be observed in the summary, the experiments are based on datasets with relatively small spatiality (up to 13 dimensions), with imbalance ratio from 9 to even 40. The

#	Dataset	Features	Samples			IR
			ALL	MAJ	MIN	
1	*ecoli-0-1-3-7-vs-2-6*	7	281	274	7	39.14
2	*ecoli4*	7	336	316	20	15.80
3	*glass-0-1-6-vs-2*	9	192	175	17	10.29
4	*glass-0-1-6-vs-5*	9	184	175	9	19.44
5	*glass2*	9	214	197	17	11.59
6	*glass4*	9	214	201	13	15.46
7	*glass5*	9	214	205	9	22.78
8	*page-blocks-1-3-vs-4*	10	472	444	28	15.86
9	*shuttle-c0-vs-c4*	9	1829	1706	123	13.87
10	*shuttle-c2-vs-c4*	9	129	123	6	20.50
11	*vowel0*	13	988	898	90	9.98
12	*yeast-0-5-6-7-9-vs-4*	8	528	477	51	9.35
13	*yeast-1-2-8-9-vs-7*	8	947	917	30	30.57
14	*yeast-1-4-5-8-vs-7*	8	693	663	30	22.10
15	*yeast-1-vs-7*	7	459	429	30	14.30
16	*yeast-2-vs-4*	8	514	463	51	9.08
17	*yeast-2-vs-8*	8	482	462	20	23.10
18	*yeast4*	8	1484	1433	51	28.10
19	*yeast5*	8	1484	1440	44	32.73
20	*yeast6*	8	1484	1449	35	41.40
21	*ecoli-0-1-4-6-vs-5*	6	280	260	20	13.00
22	*ecoli-0-1-4-7-vs-2-3-5-6*	7	336	307	29	10.59
23	*ecoli-0-1-4-7-vs-5-6*	6	332	307	25	12.28
24	*ecoli-0-1-vs-2-3-5*	7	244	220	24	9.17
25	*ecoli-0-1-vs-5*	6	240	220	20	11.00
26	*ecoli-0-2-3-4-vs-5*	7	202	182	20	9.10
27	*ecoli-0-2-6-7-vs-3-5*	7	224	202	22	9.18
28	*ecoli-0-3-4-6-vs-5*	7	205	185	20	9.25
29	*ecoli-0-3-4-7-vs-5-6*	7	257	232	25	9.28
30	*ecoli-0-3-4-vs-5*	7	200	180	20	9.00
31	*ecoli-0-4-6-vs-5*	6	203	183	20	9.15
32	*ecoli-0-6-7-vs-3-5*	7	222	200	22	9.09
33	*ecoli-0-6-7-vs-5*	6	220	200	20	10.00
34	*glass-0-1-4-6-vs-2*	9	205	188	17	11.06
35	*glass-0-1-5-vs-2*	9	172	155	17	9.12
36	*glass-0-4-vs-5*	9	92	83	9	9.22
37	*glass-0-6-vs-5*	9	108	99	9	11.00
38	*yeast-0-2-5-6-vs-3-7-8-9*	8	1004	905	99	9.14
39	*yeast-0-2-5-7-9-vs-3-6-8*	8	1004	905	99	9.14
40	*yeast-0-3-5-9-vs-7-8*	8	506	456	50	9.12

Table 1: Summary of imbalanced datasets chosen for evaluation

datasets provided by KEEL, to ensure easy comparison between results presented in various research, are already pre-divided into five parts, which forces the use of *k-fold cross-validation* with $k = 5$ in experiments (Alpaydin, 2009).

In the task of imbalanced data classification, due to its strong bias towards majority class, the *accuracy* measure is not a proper tool. For a reliable result, a measure of *balanced accuracy* is given as test results.

Both the implementation of the proposed method and the experimental environment have been constructed using the *scikit-learn* library (Pedregosa et al., 2011) in version *0.20.dev0*[1]. Among the available classification models, the MLP (*Multilayer Perceptron*) and SVC (*Support Vector Machine*) were rejected. First one was not able to build a correct model due to the lack of convergence on the small datasets (minority class of data chosen for experiments is often represented by only two patterns in cross-validated folds) and second, whose probabilistic interpretation is measurable only with sufficiently large data sets, did not allow credible construction of a fuser. As base classifiers, the following algorithms were used:

- *Gaussian Naive Bayes* (GNB) (Chan et al., 1982),
- *k-Nearest Neighbors* (kNN) — with 5 neighbors and *Minkowski* metric,
- *Decision Tree Classifier* (DTC) — with *Gini* criterion (Loh, 2011).

To provide a comparative result for the method presented in the following paper, each base classifier was also tested for the raw, imbalanced dataset and its under- and oversampled versions. Undersampling, due to high instability of results, was repeated five times on each fold. Used statistical analysis tool was a paired dependency between the classifier, which achieved the highest result and each of the others, calculated using the signed-rank *Wilcoxon* test (Wilcoxon, 1945).

Pełną implementację zaproponowanej metody i skrypt umożliwiający powtórzenie zaprezentowanych badań można odnaleźć w repozytorium abc.

The full implementation of the proposed method and the script allowing the repetition of the presented research may be found in the git repository available at `url-removed-due-to-blind-review`.

4. Experimental evaluation

Przedstawienie tabel.

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

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.

1. At the time of conducting research, only the development version of the package already has the implementation of *balanced accuracy* measure.

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	.845	.843	.843	.837	.837	.845	.845	.845	.837	.837	.841	.841	.839	.828	.825	.845	.845	.845	.837	.837	.806	.838	.825
ecoli4	.809	.834	.859	.879	.857	.748	.800	.820	.926	.898	.874	.898	.868	.913	.904	.746	.896	.895	.923	.922	.859	.787	.878
glass-0-1-6-vs-2	.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-5	.989	.989	.989	.939	.941	.989	.989	.989	.989	.989	.941	.941	.941	.941	.941	.989	.989	.989	.989	.989	.941	.975	.941
glass2	.619	.619	.619	.616	.616	.641	.641	.641	.641	.641	.619	.616	.619	.616	.616	.641	.641	.644	.641	.641	.617	.620	.591
glass4	.800	.797	.800	.802	.802	.807	.807	.804	.766	.766	.781	.781	.774	.779	.779	.812	.774	.774	.771	.771	.731	.745	.587
glass5	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.938	.945	.938
page-blocks-1-3-vs-4	.817	.817	.817	.846	.845	.828	.828	.828	.828	.828	.807	.807	.803	.846	.845	.831	.831	.831	.830	.828	.791	.806	.763
shuttle-c0-vs-c4	.994	.994	.994	.995	.994	.994	.994	.994	.991	.991	.995	.995	.995	.991	.991	.994	.994	.994	.991	.991	.990	.993	.991
shuttle-c2-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
vowel0	.909	.909	.909	.902	.903	.909	.909	.909	.909	.909	.910	.910	.910	.904	.903	.910	.909	.909	.909	.909	.906	.905	.917
yeast-0-5-6-7-9-vs-4	.690	.706	.689	.728	.731	.680	.737	.727	.786	.774	.702	.700	.677	.726	.715	.687	.724	.710	.785	.791	.498	.601	.504
yeast-1-2-8-9-vs-7	.563	.570	.567	.576	.570	.562	.557	.562	.627	.642	.563	.552	.553	.574	.566	.556	.564	.556	.625	.639	.540	.598	.544
yeast-1-4-5-8-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-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-2-vs-4	.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-8	.773	.773	.773	.773	.774	.773	.773	.773	.773	.773	.796	.771	.796	.773	.773	.773	.773	.773	.773	.773	.616	.762	.657
yeast4	.675	.674	.674	.769	.728	.708	.766	.766	.822	.813	.644	.655	.637	.781	.746	.710	.765	.769	.820	.811	.526	.660	.551
yeast5	.934	.934	.934	.934	.934	.929	.934	.938	.957	.955	.917	.921	.919	.934	.934	.927	.930	.935	.957	.955	.780	.910	.831
yeast6	.785	.787	.787	.847	.796	.829	.863	.848	.887	.878	.760	.773	.768	.843	.795	.818	.860	.845	.887	.878	.628	.795	.650
ecoli-0-1-4-6-vs-5	.785	.781	.783	.804	.804	.660	.810	.810	.906	.913	.810	.812	.860	.906	.906	.785	.837	.862	.910	.913	.885	.679	.877
ecoli-0-1-4-7-vs-2-3-5-6	.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-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-vs-2-3-5	.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-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-0-2-3-4-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-0-2-6-7-vs-3-5	.563	.563	.588	.635	.613	.588	.588	.588	.625	.630	.588	.588	.588	.625	.628	.588	.613	.613	.623	.628	.592	.595	.563
ecoli-0-3-4-6-vs-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-0-3-4-7-vs-5-6	.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-vs-5	.756	.756	.756	.833	.792	.711	.786	.783	.883	.883	.769	.794	.811	.772	.758	.758	.783	.783	.883	.883	.730	.657	.817
ecoli-0-4-6-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-0-6-7-vs-3-5	.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-5	.795	.787	.793	.805	.812	.685	.755	.705	.853	.830	.845	.838	.838	.828	.825	.688	.755	.755	.853	.830	.847	.682	.780
glass-0-1-4-6-vs-2	.558	.589	.558	.615	.581	.622	.622	.620	.620	.620	.592	.592	.558	.595	.595	.622	.622	.620	.620	.620	.597	.590	.577
glass-0-1-5-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-0-4-vs-5	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.994	.984	.984	.994
glass-0-6-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
yeast-0-2-5-6-vs-3-7-8-9	.576	.576	.576	.576	.576	.576	.576	.576	.782	.782	.634	.657	.692	.777	.768	.576	.576	.576	.781	.783	.782	.605	.670
yeast-0-2-5-7-9-vs-3-6-8	.883	.893	.889	.901	.901	.901	.901	.902	.897	.900	.779	.834	.841	.900	.898	.894	.896	.896	.897	.900	.524	.785	.577
yeast-0-3-5-9-vs-7-8	.597	.632	.606	.620	.621	.606	.605	.606	.605	.605	.601	.607	.589	.619	.633	.611	.612	.600	.605	.605	.539	.633	.557

Table 2: Balanced accuracy scores obtained using GNB 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				
.836	.836	.836	.847	.849	.845	.845	.847	.844	.845	.830	.830	.832	.835	.834	.854	.853	.851	.853	.856	.835	.835	.850	*ecoli-0-1-3-7-vs-2-6*
.941	.941	.945	.949	.949	.940	.940	.940	.921	.945	.964	.967	.972	.909	.909	.945	.945	.945	.940	.943	.909	.928	.848	*ecoli4*
.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-2*
.880	.880	.880	.883	.883	.830	.877	.880	.880	.880	.836	.839	.853	.879	.879	.833	.833	.836	.833	.841	.933	.852	.739	*glass-0-1-6-vs-5*
.696	.696	.724	.721	.719	.701	.724	.718	.724	.724	.737	.757	.746	.756	.756	.726	.732	.731	.723	.746	.715	.678	.485	*glass2*
.861	.861	.861	.878	.875	.861	.861	.861	.863	.863	.868	.868	.888	.905	.913	.861	.861	.861	.888	.898	.925	.865	.781	*glass4*
.823	.823	.826	.830	.830	.813	.813	.813	.813	.816	.838	.843	.879	.801	.873	.816	.813	.826	.821	.833	.830	.811	.695	*glass5*
.868	.868	.874	.877	.877	.867	.871	.866	.902	.897	.901	.907	.926	.909	.924	.874	.875	.881	.917	.920	.917	.872	.808	*page-blocks-1-3-vs-4*
.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	.996	*shuttle-c0-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	*shuttle-c2-vs-c4*
.942	.942	.942	.944	.943	.943	.942	.943	.945	.945	.952	.953	.963	.997	.996	.948	.948	.954	.980	.986	.999	.939	.977	*vowel0*
.781	.781	.785	.786	.790	.779	.779	.785	.785	.787	.810	.824	.832	.830	.820	.799	.800	.819	.828	.829	.795	.792	.667	*yeast-0-5-6-7-9-vs-4*
.644	.645	.662	.665	.663	.659	.660	.655	.645	.659	.702	.695	.673	.625	.626	.655	.660	.684	.692	.693	.627	.652	.499	*yeast-1-2-8-9-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-1-4-5-8-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-1-vs-7*
.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-4*
.715	.715	.740	.719	.719	.721	.721	.728	.727	.734	.782	.782	.778	.803	.803	.738	.738	.752	.755	.758	.803	.734	.774	*yeast-2-vs-8*
.819	.819	.819	.839	.838	.840	.841	.841	.841	.843	.838	.829	.839	.771	.774	.844	.844	.843	.841	.852	.749	.835	.574	*yeast4*
.955	.955	.954	.955	.954	.957	.957	.956	.957	.958	.956	.956	.960	.962	.967	.958	.958	.958	.960	.960	.964	.952	.850	*yeast5*
.883	.883	.884	.886	.885	.886	.886	.885	.887	.886	.899	.900	.892	.840	.865	.887	.887	.888	.893	.895	.840	.879	.739	*yeast6*
.890	.890	.890	.892	.892	.892	.892	.892	.892	.890	.894	.894	.896	.904	.906	.892	.892	.892	.892	.892	.917	.886	.898	*ecoli-0-1-4-6-vs-5*
.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-2-3-5-6*
.886	.886	.888	.868	.868	.883	.885	.881	.913	.891	.874	.876	.887	.892	.894	.893	.893	.871	.887	.889	.899	.883	.838	*ecoli-0-1-4-7-vs-5-6*
.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-2-3-5*
.905	.905	.907	.914	.911	.905	.905	.907	.911	.911	.911	.911	.911	.911	.914	.909	.909	.911	.911	.911	.916	.902	.900	*ecoli-0-1-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-0-2-3-4-vs-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-0-2-6-7-vs-3-5*
.887	.887	.874	.890	.890	.887	.887	.879	.890	.890	.893	.893	.895	.901	.901	.893	.893	.890	.895	.895	.911	.881	.875	*ecoli-0-3-4-6-vs-5*
.904	.904	.902	.917	.917	.904	.904	.902	.921	.915	.901	.901	.901	.890	.890	.913	.911	.897	.885	.888	.894	.887	.876	*ecoli-0-3-4-7-vs-5-6*
.903	.903	.903	.903	.900	.900	.900	.903	.903	.900	.906	.906	.906	.903	.903	.906	.906	.906	.903	.903	.911	.889	.875	*ecoli-0-3-4-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-0-4-6-vs-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-3-5*
.853	.853	.850	.868	.863	.860	.860	.858	.870	.872	.878	.878	.880	.863	.885	.865	.865	.872	.890	.887	.863	.850	.847	*ecoli-0-6-7-vs-5*
.715	.715	.706	.712	.725	.717	.712	.712	.714	.717	.731	.717	.701	.718	.716	.741	.749	.747	.734	.727	.732	.681	.512	*glass-0-1-4-6-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-0-1-5-vs-2*
.932	.932	.939	.932	.932	.939	.939	.944	.939	.944	.951	.951	.951	.951	.951	.951	.951	.951	.951	.951	.988	.917	.850	*glass-0-4-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	*glass-0-6-vs-5*
.784	.784	.786	.781	.784	.785	.785	.790	.786	.790	.801	.802	.800	.791	.798	.793	.798	.800	.795	.798	.784	.760	.762	*yeast-0-2-5-6-vs-3-7-8-9*
.909	.909	.908	.911	.911	.904	.904	.908	.906	.910	.913	.913	.902	.900	.900	.909	.913	.912	.896	.895	.904	.902	.902	*yeast-0-2-5-7-9-vs-3-6-8*
.757	.757	.755	.746	.753	.748	.748	.757	.749	.754	.759	.742	.739	.731	.730	.770	.767	.750	.723	.734	.718	.702	.639	*yeast-0-3-5-9-vs-7-8*

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

Without oversampled set: All members REG	WEI	CON	NOR	NC	Without oversampled set: Reduced members REG	WEI	CON	NOR	NC	With oversampled set: All members REG	WEI	CON	NOR	NC	With oversampled set: Reduced members REG	WEI	CON	NOR	NC	OS	US	Full	Dataset
.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 4: Balanced accuracy scores obtained using DTC as a base classifier

References

Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, and Francisco Herrera. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011.

Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *Pattern recognition (ICPR), 2010 20th international conference on*, pages 3121–3124. IEEE, 2010.

T. F. Chan, G. H. Golub, and R. J. LeVeque. Updating formulae and a pairwise algorithm for computing sample variances. In H. Caussinus, P. Ettinger, and R. Tomassone, editors, *COMPSTAT 1982 5th Symposium held at Toulouse 1982*, pages 30–41, Heidelberg, 1982. Physica-Verlag HD. ISBN 978-3-642-51461-6.

Alberto Fernández, María José del Jesus, and Francisco Herrera. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *International Journal of Approximate Reasoning*, 50(3):561–577, 2009.

Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6): 80–83, 1945.