Dealing with the task of imbalanced, multidimensional data classification using ensembles of *exposers*

Paweł Ksieniewicz PAWEL.KSIENIEWICZ@PWR.EDU.PL
Michał Woźniak MICHAL.WOZNIAK@PWR.EDU.PL
Department of Systems and Computer Networks
Faculty of Electronics
Wroclaw University of Science and Technology

Editors: Luís Torgo, Bartosz Krawczyk, Paula Branco and Nuno Moniz.

Abstract

Recently, the problem of imbalanced data is the focus of intense research of machine learning community. Following work tries to utilize an approach of transforming the data space into another where classification task may become easier. Paper contains a proposition of a tool, based on a photographic metaphor to build a classifier ensemble, combined with a *random subspace* approach. Developed solution is insensitive to a sample size and robust to dimension increase, which allows a regularization of feature space, reducing the impact of biased classes. The proposed approach was evaluated on the basis of the computer experiments carried out on the benchmark and synthetic datasets.

Keywords: curse of dimensionality, imbalance data, random subspace, ensemble classifiers

1. Introduction

If the classes in available dataset are not represented equally, we are encountering a problem of *imbalanced data*. In extreme intensity of this situation, being relatively common, i.e., in the fraud detection problems (Phua et al., 2004) or in the medical data analysis (Mazurowski et al., 2008), where usually occurs a minority report on most important class (a fraud or a sickness), often happens the *accuracy paradox* (Valverde-Albacete and Peláez-Moreno, 2014). It means that for a strongly uneven distribution of classes, higher *accuracy*, being the most common measure of classifier performance in literature (Demsar, 2006), does not indicate the greater discriminative power.

Two most popular approaches (Krawczyk, 2016) to deal with problems caused by *imbalanced data* are (i) *inbuilt mechanisms*, which change the classification rules to enforce a bias toward the minority class using e.g., cost-sensitive approach (Krawczyk et al., 2014) and (ii) *data preprocessing methods*, which modify the data distribution to change the balance between classes. The preprocessing approach uses *oversampling* of a minority class, applied usually when overall number of objects in dataset is relatively low (He and Garcia, 2009) and *undersampling* of a majority class (Japkowicz and Stephen, 2002), popular for large datasets (Liu et al., 2009). Both of them have its flaws. *Oversampling*, especially conducted before *cross-validation* comes with a risk of overfitting, while *undersampling* may lead to significant decrease of classifiers discrimination power due to removal of informative objects.

Despite the risk of *overtraining* and problems to measure the classifiers quality in a cases of *imbalanced datasets*, one of the biggest problems in *pattern recognition* is the *curse of dimensionality* (Bellman, 1961). According to it, the growth of feature vector, causes that the generalizations are becoming exponentially harder. So it is necessary not to forget about the structure of single object itself. While it is a vector of d dimensions, not all of them are equally relevant for *pattern recognition* methods. Some of them, by injecting nothing more than a noise, may have only negative aspect on the quality of our algorithms (Segura et al., 2003). Some of them may turn out to be exploitable only in relation to others (Bishop, 2006). Having all this information in mind, it is important to know about dimensionality reduction techniques.

The *feature selection* aims to choose a subset of original feature set, without changing its values. While it was proved as an useful and effective technique (Liu and Motoda, 2009), it leaves original features untouched, remaining their physical meanings to be still interpretable by a human being (Li and Liu, 2016). This advantage encourages to use in real world applications (Kursa and Rudnicki, 2010), because it causes a loss only of data irrelevant for classification (Rudnicki et al., 2015). The second technique, *feature extraction* (Stapor, 2011), tries to create a function to map d dimensional vector into smaller s dimensional one. Even since we reuse information from even a whole vector, original features are here replaced by their scalar products and we are no longer able to interpret them directly.

While feature selection is rather a discrete process, the optimization of feature extraction methods is quite *smoother* and often leads for a more fitting solution (Koller and Sahami, 1996), by employing information from wider range than only resulting subspace and precisely setting the impact of original features onto the transformation result.

On the other hand, there is no need to hold down only one algorithm into a single problem. We can also compose structures of a multiple classifier system, which combines a set of classifiers to provide a common decision of the *classifier ensemble* (Kuncheva, 2014). The main element of this structure is a *pool of classifiers*. The most important aim of good selection of the member classifiers, is to provide their high diversity, which means that each classifier should make an independent decision (Dietterich, 2000). We can try to ensure it by differencing the input and output data, or by differencing the classifier models. One of the popular methods, is a *random subspace* approach, originally implemented for *decision trees*, which brought the *random forests* (Ho, 1998) and later applied with success also for SVM*'s*, *linear classifiers* or k-NN. It tries to weaken the correlation of classifiers in *ensemble*, by letting each of them learn on a reduced, random subset of features, introducing the approach of random *feature selection*.

The main contributions of this work are:

- the proposition of the exposer – a visualization tool for numerical data distribution, usable as *Exposer Classifier* – the *supervised learning* method to use it in a task of decision making, being the component to build *Exposer Classifier Ensemble* – the *Random Subspace* based approach of classifier ensemble.

- an experimental evaluation of the proposed concept presenting the detailed results that offers an in-depth insight into the importance of selecting proper examples for the oversampling procedure. A dedicated website[1] presents detailed results.

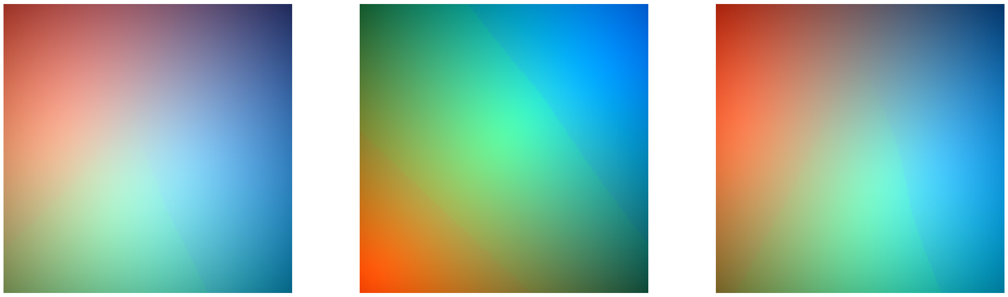
[1] http://ksienie.com/ecml17/

- a set of conclusions that will allow to design efficient classifiers for datasets.

2. Exposers

Following section aims to explain the idea of *exposer*, a data representation, based on visualization tools of numerical data distribution, creating, from a regular dataset, the spacial-spectral structure similar to a multi-spectral image (Ksieniewicz et al., 2017). However, *exposer* is not intended to work as a preprocessing tool, but as a classifier itself. Moreover, composing a set of *exposers*, generated on different feature subsets of a same training set, leads to *Exposer Classifier Ensemble* (ECE).

A proposed data structure is drawing from tenets of *histogram* and a *scatter plot*. Like in *histogram*, the range of values is divided into a series of intervals, but like in a *scatter plot*, not the single value, but the combination of them is analyzed. The rule of adjacency is broken here, so each object may fall into more than one of the bins. Exemplary *exposers*, prepared for two-dimensional feature subsets of *iris* dataset are presented in Figure 1.



(*a*) 1:2

(*b*) 3:4

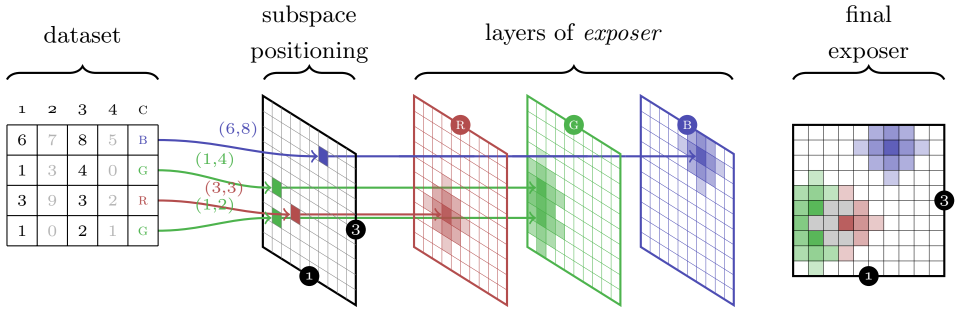
(*c*) 2:3

Figure 1: Exemplary exposers for *iris* dataset. Digits points the indexes of features.

The procedure of creating an *exposer* is inspired by the process of plate light exposure in the chemical photography. Hence, its control parameters are plate grain (*exposer* counterpart of *histograms* bin) and a light dispersion factor (named a *radius*, as a relative width of a bin according to a range of values). Instead of exposing the photographic plate coated with light-sensitive chemicals to the light source, a numerical representation matrix is exposed to the beams projected from the data samples. Procedure *takes a photography* of a subset of features of the data samples, where intensity of a *light* in every point is a density aggregation of the data samples falling in its neighborhood.

Exemplary process of exposing is illustrated in Figure 2. We have a dataset which consists of four objects, each described by four features[2] and assigned to one of three classes (R, G and B). The grain parameter is fixed at 10 and selected subspace uses pair of first and third feature. At the first step, *subspace positioning*, every object in dataset is placed on a two-dimensional grid, divided by 10 in every dimension, according to values of chosen features. The exposer consists of three layers, one per every class in dataset, and every layer

[2] To clarify illustration, all the feature values in exemplary dataset are integers in range 0–10. Real implementation normalizes original values as floats in range 0–1.



1	2	3	4	C
6	7	8	5	B
1	3	4	0	G
3	9	3	2	R
1	0	2	1	G

Figure 2: Exposing four samples covering three classes on two-dimensional subspace (1,3).

is influenced only by objects labeled accordingly, by adding the negative distance (where closer means higher value) between positioned centre of object to corresponding cell.

The most important difference between classic photography and *exposers* is a redefinition of the concept of color, being not the classical RGB vector nor a real *spectral signature*. For *exposers* it consists not of classical three spectral channels (Svaetichin, 1956), but of one dimension per class of the dataset. Hence, the representation matrix has as many layers as classes. The representation matrix exposure process sensitizes each layer projecting to it only objects from the corresponding class. There may be assumed, that *exposing* procedure generates some kind of multispectral imaging from the data, which counterpart of *spectral signature* is interpreted as a *support vector* during the classification procedure.

To prepare a mathematical description of *exposer*, let $\mathcal{DS}$ denote a set of n examples, where each of them x_k is represented by the d-dimensional feature vector and its label i_k from the finite set of available labels $\mathcal{M}$.

$$\begin{array}{lcll}
\mathcal{X} & \subseteq & \Re^d & \\
\mathcal{M} & = & \{1, 2, \ldots, M\} & \\
x_k & = & [x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(d)}], & x_k \in \mathcal{X} \\
i_k & = & [i_1, i_2, \ldots, i_n], & i_k \in \mathcal{M} \\
\mathcal{DS} & = & \Big\{(x_1, i_1), (x_2, i_2), \ldots, (x_n, i_n)\Big\} &
\end{array} \tag{1}$$

Variety of possible exposers comes from a set Λ of $\binom{d}{s}$ combinations λ_i, where s is the chosen exposer dimension, i.e., number of chosen features.

$$\begin{array}{lcll}
\Lambda & = & \{\lambda_1, \lambda_2, \ldots, \lambda_L\}, & |\Lambda| = L = \binom{d}{s} \\
\lambda_i & = & [l_1, l_2, \ldots, , l_s], & l_j \in \{1, 2, \ldots, d\}, \quad l_1 \neq l_2 \neq \ldots \neq l_s,
\end{array} \tag{2}$$

Representation of *exposer* ($\mathcal{E}$) is a s-dimensional cube

$$\begin{aligned}\mathcal{E}_m &\in G^s = \underbrace{G \times G \times \ldots \times G}_{s} \\ \mathcal{E} &= \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_M\}.\end{aligned} \tag{3}$$

And every point positioned by loc gives us a vector of M values v

$$\begin{aligned}\mathcal{E}^{(loc)} &= [v_1, v_2, \ldots, v_M]^T \\ loc &= [loc_1, loc_2, \ldots, loc_s]^T\end{aligned} \tag{4}$$

A single value of *exposer* is a sum of all positive differences between a given radius r and a distance from a real vector of grid cell central point loc to the location loc_k of samples in a subspace, labeled accordingly to the mth layer.

$$\begin{aligned}\mathcal{E}_m^{(loc)} &= \textstyle\sum_{k=1}^{n} \Big[d(loc, loc_k) < r \ \wedge \ i_k = m\Big] \cdot \Big(r - d(loc, loc_k)\Big) \\ loc_k &= [x_k^{(\lambda_1)}, x_k^{(\lambda_2)}, \ldots, x_k^{(\lambda_s)}]^T\end{aligned} \tag{5}$$

which is a discretization on the g quants in every spacial dimension created by given λ_i.

According to the photographic metaphor, the previous expression may be imagined as a projection by exposure, where every x_k sample is the *photon* with location described with features chosen by combination λ, affecting the image in r radius.

For the classification task, there is an *exposer* $\mathcal{E}$, exposed using a subspace λ of a *learning set* $\mathcal{LS}$ and a testing sample x_k from a *testing set* $\mathcal{TS}$ to classify.

$$x_k = [x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(d)}]^T, \qquad x_k \in \mathcal{TS} \tag{6}$$

Classifier Ψ makes a decision on the basis of the class index with the maximum value from a value vector corresponding to the testing sample.

$$\Psi(x_k) = \underset{m \in \mathcal{M}}{argmax} \left(\mathcal{E}_m^{(loc_k)}\right) \tag{7}$$

Increase of the feature vector also rapidly increases a number of possible *exposers* to obtain. It creates a possibility to gather a set of *exposer* classifiers into an ensemble. To establish such ensemble, a prediction procedure needs a little enhancement. There is a testing sample x_k to classify, but this time we have an ensemble of *exposers* (Π) build around the set of combinations Λ', being a subset of all possible combinations

$$\begin{aligned}\Lambda' &\subset \Lambda, & |\Lambda'| = N, \qquad N < L \\ \Pi &= \{\Psi_1, \Psi_2, \ldots, \Psi_N\}, & \Psi : \mathcal{X} \leftarrow \mathcal{M}\end{aligned} \tag{8}$$

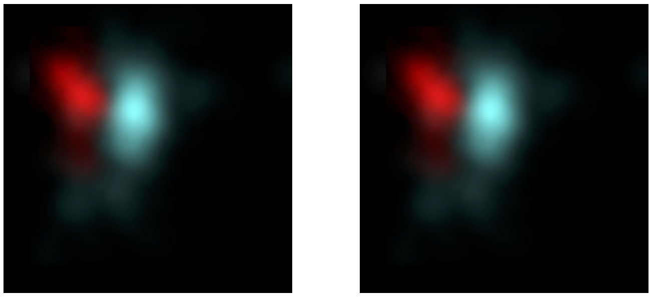
Exposer may be visualized as a regular RGB substitution if number of classes, like in example from Figure 2, matches the number of color channels observable by eyes. To make a color visualization for number of classes other than 3, the *hue, saturation, value* (HSV) (Smith, 1978) interpretation of *exposer* point was proposed. The value (V) is the maximum of $\mathcal{E}^{(loc)}$ vector and the *hue* (H) is an angle of its index normalized to 360°.

$$\begin{aligned} V(\mathcal{E}^{(loc)}) &= max(\mathcal{E}^{(loc)}) \\ H(\mathcal{E}^{(loc)}) &= \underset{i \in \mathcal{M}}{argmax}(\mathcal{E}^{(loc)}) \cdot \tfrac{360^\circ}{M} \end{aligned} \tag{9}$$

Although the vector $\mathcal{E}^{(loc)}$, like a spectral signature of multi-spectral data, does not point a specific color of a visible spectrum, we can still interpret the grey concealed in it in the same way as in the classic color theory, meaning an equal presence of any value building it. Black corresponds to values close to nothing, while white stands for the maximum values. According to it, *saturation* is defined as a difference between the minimum and maximum of the $\mathcal{E}^{(loc)}$.

$$S(\mathcal{E}^{(loc)}) = max(\mathcal{E}^{(loc)}) - min(\mathcal{E}^{(loc)}) \tag{10}$$

Figure 3 presents exemplary *exposers*, with relatively small radius parameter, based on the chosen feature pairs, acquired for the *yeast3* dataset. The dataset consists of 1483 samples described by 8 features, divided into two classes. The visualization aggregates both layers into one by colour coding each class (*red* – positive, *green* – negative) modulated by the layer intensity.



(a) 1:3

(b) 1:4

Figure 3: *Exposers* created using two example combinations for *yeast3* dataset.

The left image presents a two feature subspace, where colors are highly mixing. There is a high presence of grays, and *class colors* are faded, so intuitively, this pair of features should provide a weak discrimination power. The image on the right side shows a two feature subspace where mixing of colors is minimal, so it may be assumed, that classes are highly separable in this subspace.

Finally, a three level classifier ensemble is proposed, which idea is depicted in Figure 4. At the lowest level there is a set of monochrome layers, each of which defines a member classifier, characterized by a combination of features, denoted by λ_i, and a weight Θ_i used to combine its output with the remaining classifiers into an *exposer*. At the top level, member classifiers are combined into an ECE ensemble.

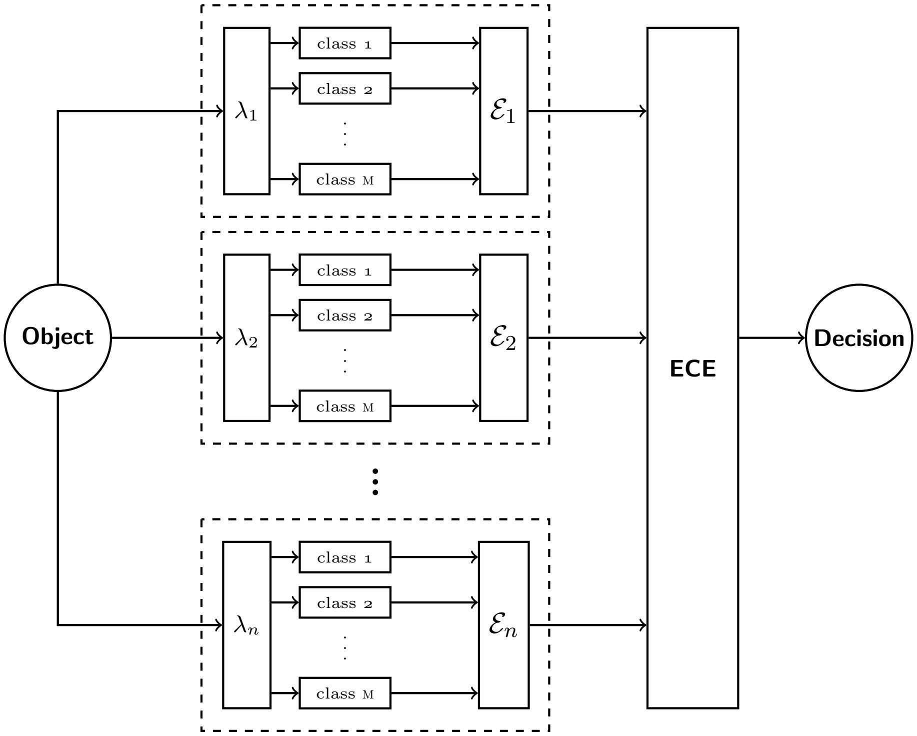


Figure 4: ECE classification diagram

3. Experimental evaluation

Set-up The experimental evaluation of proposed method was realized under *Weles* framework (Ksieniewicz, 2017). Its full code is available to access as a GIT repository on *Github* page of *Department of Systems and Computer Networks*[3] and published in in *Python Package Index*. Quality of all the written *Python* code is statically analyzed using *Code Climate* and is continuously integrated using *Travis* CI. Current quality metrics and test coverages are listed in the homepage of every repository.

For the evaluation of algorithm capabilities, eight benchmark datasets were used. All of them contains imbalanced binary problems. Four of them are coming from UCIMLR (Lichman, 2013) and four are synthetic, generated data with various number of features (from 20 to 200) and fixed 9:1 imbalance ratio. Also five reference classification algorithms were selected. To use them without a need of a new implementation, *Weles* framework was enhanced by the adapter able to wrap classifiers from a popular machine learning library *scikit-learn* (Pedregosa et al., 2011), to encapsulate prepared selection in a separate classes.

[3] *https://github.com/w4k2*

The classifiers were configured with most popular parameters, to provide a standard, *state-of-art* pool for proper comparison of proposed method. Prepared datasets and classifiers pool were cross-tested using measure of *balanced accuracy*(Brodersen et al., 2010) (Table 1).

Table 1: Balanced accuracy achieved by chosen reference algorithms and data characteristics.

Dataset	KNN	GNB	DTC	MLP	SVC	Characteristics		
	Minimal Distance Classifier	Bayesian Classifier	Rule-Based Classifier	Artificial Neural Network	Support Vector Machine	Samples	Features	Imbalance ratio
balance	0.768	0.786	0.720	0.635	0.785	625	4	6:1
ionosphere	0.692	0.845	0.795	0.592	0.590	351	34	2:1
wisconsin	0.943	0.954	0.898	0.500	0.946	699	9	2:1
yeast3	0.814	0.575	0.811	0.858	0.500	1484	8	8:1
synthetic20	0.565	0.768	0.702	0.707	0.500	4000	20	9:1
synthetic50	0.548	0.765	0.867	0.500	0.500	4000	50	9:1
synthetic100	0.561	0.767	0.788	0.789	0.500	4000	100	9:1
synthetic200	0.511	0.623	0.881	0.500	0.500	4000	200	9:1

Experiment The experiments were conducted under 5×2 *cross-validation* (Alpaydin, 2014) and presented results are a BAC value. Red color means that the result is statistically significantly better than results presented in black, and statistical significancy was measured using T-test. Last row of result tables will contain the result obtained by the best and the worst reference classifier, extended by a difference between score of ECE and leading reference solution.

Table 2: Experiment configuration table

Fixed parameters	
dimensions	2
approach	random
limit	30
Studied parameters	
grain	4 − 32
radius	.1 − .5

As a test of ECE, it is necessary to establish the values of parameters of single *exposer* configuration: its *radius* and *grain*, both of which being exponential reason of growth of the computational time needed to process a training set. For datasets with feature vector larger than 8 values, there is no chance to obtain *exposers* for all possible subspaces in a reasonable time. To deal with this problem, a *random subspace* approach was used. The planned experiment uses eight benchmark datasets. For a low computational complexity, only two-dimensional subspaces are considered.

Discussion Comparison of balanced accuracy between ECE and reference classifiers for imbalanced binary problems is presented in Table 3. It is worth observing that the ECE

Table 3: Balanced accuracy achieved on all tested datasets using ECE.

(*a*) balance

Radius	Grain			
	4	8	16	32
0.1	0.500	0.500	0.653	0.699
0.2	0.500	0.653	0.699	0.701
0.3	0.690	0.705	0.704	0.720
0.4	0.690	0.730	0.735	0.736
0.5	0.724	0.732	0.730	0.728

-5.0%, best 0.786 (GNB), worst 0.635 (MLP)

(*b*) ionosphere

Radius	Grain			
	4	8	16	32
0.1	0.500	0.500	0.500	0.863
0.2	0.500	0.763	0.842	0.859
0.3	0.500	0.794	0.863	0.878
0.4	0.500	0.824	0.797	0.840
0.5	0.710	0.740	0.757	0.752

+3.3%, best 0.845 (GNB), worst 0.590 (SVC)

(*c*) wisconsin

0.1	0.500	0.500	0.927	0.957
0.2	0.500	0.955	0.965	0.964
0.3	0.959	0.968	0.969	0.969
0.4	0.958	0.967	0.968	0.970
0.5	0.967	0.966	0.970	0.970

+1.6%, best 0.954 (GNB), worst 0.500 (MLP)

(*d*) yeast3

0.1	0.500	0.500	0.708	0.817
0.2	0.500	0.771	0.869	0.881
0.3	0.832	0.840	0.873	0.880
0.4	0.830	0.848	0.874	0.883
0.5	0.852	0.855	0.884	0.889

+3.1%, best 0.858 (MLP), worst 0.500 (SVC)

(*e*) synthetic20

0.1	0.500	0.500	0.548	0.527
0.2	0.500	0.583	0.696	0.613
0.3	0.642	0.650	0.760	0.818
0.4	0.744	0.839	0.839	0.880
0.5	0.764	0.816	0.838	0.840

+11.2%, best 0.768 (GNB), worst 0.500 (SVC)

(*f*) synthetic50

0.1	0.500	0.500	0.512	0.503
0.2	0.500	0.530	0.615	0.687
0.3	0.583	0.508	0.771	0.699
0.4	0.661	0.802	0.833	0.732
0.5	0.677	0.798	0.764	0.887

+2.0%, best 0.867 (DTC), worst 0.500 (MLP)

(*g*) synthetic100

0.1	0.500	0.500	0.507	0.500
0.2	0.500	0.537	0.621	0.577
0.3	0.560	0.827	0.620	0.868
0.4	0.521	0.843	0.593	0.513
0.5	0.611	0.739	0.873	0.890

+10.1%, best 0.789 (MLP), worst 0.500 (SVC)

(*h*) synthetic200

0.1	0.500	0.500	0.501	0.500
0.2	0.500	0.500	0.510	0.501
0.3	0.533	0.529	0.504	0.662
0.4	0.506	0.661	0.694	0.494
0.5	0.532	0.635	0.602	0.816

-6.5%, best 0.881 (DTC), worst 0.500 (MLP) often achieves its high effectiveness on relatively low values of tested parameters (.3 radius and grain of 32). However, the hardest, high-dimensional and highly imbalanced problems of *synthetic*20 and *synthetic*50 shows, that further growth of grain parameter may have positive impact on classification.

In most of cases (the only exception is *balance* dataset), proposed solution outperforms all of reference methods, in one case even by over 10%. It is never the worst solution in the competitive pool.

The algorithm, by employing the information of features distribution, takes all the advantages from the characteristics of algorithms similar to *bayesian classifiers*, achieving the best results with imbalanced data from the reference classifiers pool. Moreover it remains immune to high-dimensional data due to *random subspace* approach. Connecting it with method of making decisions typical to *minimal distance classifiers*, occurs to be a solid solution for problematic cases of both imbalanced and high-dimensional datasets.

4. Conclusions

Following paper presented a classifier, formulated as an ensemble of subspace projections spanned on combined features of data. In the tested approach, each image corresponds to a feature pair subspaces.

ECE allows a dynamical feature extraction, adjusting itself to every tested object. The fitting procedure randomly selects *attributes* using *random subspace* method and *feature extraction* is done at the classification stage, which makes it a two-level *feature reduction* and *classification* method. All the attributes are fully interpretable, but invertible, taking the best advantages of both *feature selection* and *extraction* methods. It is possible due to discretization of the *exposer* feature space.

It was shown, that this approach led to create classifier that is competitive to existing ones and able to outperform them for some kind of data, proving that it can be used for real-life applications. The method shows robustness to the *curse of dimensionality*, which allows processing the big data problems and a high performance with imbalanced problems.

Acknowledgments

This work was supported by the Polish National Science Center under the grant no. UMO-2015/19/B/ST6/01597 and by the statutory fund of the Faculty of Electronics, Wroclaw University of Science and Technology.

References

Ethem Alpaydin. *Introduction to Machine Learning.* MIT Press, August 2014. ISBN 0262028182.

Richard Bellman. On the reduction of dimensionality for classes of dynamic programming processes. *Journal of Mathematical Analysis and Applications*, 3(2):358–360, October 1961. doi: 10.1016/0022-247X(61)90062-2.

Christopher Bishop. *Pattern Recognition and Machine Learning.* Springer-Verlag New York, 2006. ISBN 978-0-387-31073-2.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition (ICPR)*, pages 3121–3124. IEEE, 2010. ISBN 978-1-4244-7542-1. doi: 10.1109/ICPR.2010.764.

Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, January 2006.

Thomas G Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15. Springer, Berlin, Heidelberg, Berlin, Heidelberg, June 2000. ISBN 978-3-540-67704-8. doi: 10.1007/3-540-45014-9_1.

H He and E A Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21:1263–1284, September 2009. doi: 10.1109/TKDE.2008.239.

T K Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine ...*, 1998. doi: 10.1109/34.709601.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002.

Daphne Koller and Mehran Sahami. Toward Optimal Feature Selection. In *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996*, pages 284–292, 1996.

B. Krawczyk, M. Woźniak, and G. Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554–562, 2014.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, Nov 2016. ISSN 2192-6360. doi: 10.1007/s13748-016-0094-0.

Paweł Ksieniewicz. Weles - A Machine Learning library as simple as medieval village. *GitHub repository, https://github.com/w4k2/weles*, 2017.

Paweł Ksieniewicz, Bartosz Krawczyk, and Michal Wozniak. Ensemble of Extreme Learning Machines with Trained Classifier Combination and Statistical Features for Hyperspectral Data. *Neurocomputing*, pages –, 2017. doi: https://doi.org/10.1016/j.neucom.2016.04.076.

L.I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms.* Wiley, 2014. ISBN 9781118914557.

Miron B Kursa and Witold R Rudnicki. Feature Selection with the Boruta Package . *Journal of Statistical Software*, 36(11):1–13, September 2010. doi: 10.18637/jss.v036.i11.

Jundong Li and Huan Liu. Challenges of Feature Selection for Big Data Analytics. *arXiv.org*, November 2016.

M Lichman. UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science., 2013.

Huan Liu and Hiroshi Motoda. *Computational Methods of Feature Selection*, volume 45. Inf. Process. Manage., Boca Raton, FL, 2009. ISBN ISBN 978-1-58488-878-9. doi: 10.1016/j.ipm.2009.03.003.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, April 2009. doi: 10.1109/TSMCB.2008.2007853.

Maciej A Mazurowski, Piotr A Habas, Jacek M Zurada, Joseph Y Lo, Jay A Baker, and Georgia D Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2-3): 427–436, March 2008. doi: 10.1016/j.neunet.2007.12.031.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter*, 6(1):50–59, June 2004. doi: 10.1145/1007730.1007738.

Witold R Rudnicki, Mariusz Wrzesien, and Wieslaw Paja. All Relevant Feature Selection Methods and Applications. *Feature Selection for Data and Pattern Recognition*, 584 (Chapter 2):11–28, 2015. doi: 10.1007/978-3-662-45620-0_2.

José C Segura, Javier Ramirez, M Carmen Benitez, Ángel de la Torre, and Antonio J Rubio. Improved feature extraction based on spectral noise reduction and nonlinear feature normalization. In *EUROSPEECH-2003*, pages 353–356, September 2003.

Alvy Ray Smith. Color gamut transform pairs. *ACM SIGGRAPH Computer Graphics*, 12 (3):12–19, August 1978. doi: 10.1145/965139.807361.

Katarzyna Stapor. *Metody klasyfikacji obiektów w wizji komputerowej*. Wydawnictwo Naukowe PWN, 2011. ISBN 978-8-3011-6581-9, 9788301165819.

G Svaetichin. Spectral response curves from single cones. *Acta Physiologica Scandinavica*, 39(134):17–46, 1956.

Francisco J Valverde-Albacete and Carmen Peláez-Moreno. 100Information Transfer Factor Explains the Accuracy Paradox. *PLOS ONE*, 9(1):e84217, January 2014. doi: 10.1371/journal.pone.0084217.