Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 247–251
DOI: 10.15439/2023F3838
ISSN 2300-5963 ACSIS, Vol. 35

Towards automated detection of adversarial attacks on tabular data

Piotr Biczyk*§, Łukasz Wawrowski†
* Silesian University of Technology, Faculty of Automatic Control, Electronics and Computer Science
Akademicka 16, 44-100 Gliwice, Poland, pbiczyk@gmail.com
† Łukasiewicz Research Network, Institute of Innovative Technologies EMAG
ul. Leopolda 31, 40-189 Katowice, Poland, lukasz.wawrowski@emag.lukasiewicz.gov.pl
§ QED Software sp. z o.o., Mazowiecka 11/49, 00-052 Warsaw, Poland

***Abstract*—The paper presents a novel approach to investigating adversarial attacks on machine learning classification models operating on tabular data. The employed method involves using diagnostic parameters calculated on an approximated representation of a model under attack and analyzing differences in these diagnostic parameters over time. The hypothesis researched by the authors is that adversarial attack techniques, even if attempting a low-profile modification of input data, influence those diagnostic attributes in a statistically significant way. Thus, changes in diagnostic attributes can be used for detecting attack events. Three attack approaches on real-world datasets were investigated. The experiments confirm the approach as a promising technique to be further developed for detecting adversarial attacks.**

I. Introduction

THE widespread adoption of machine learning (ML) algorithms in various fields, such as healthcare, finance, transportation, and industry [1], has revolutionized the way we process and analyze vast amounts of data [2]. However, the rapid proliferation of ML applications has also raised operational security concerns, as malicious actors increasingly target these models with adversarial attacks to undermine their reliability and compromise their performance [3]. These attacks pose a significant threat to the integrity and trustworthiness of ML models, necessitating the development of robust detection and mitigation techniques to protect the systems from potential threats [4], [5].

The motivation for our work is rooted in the observed disparity between machine learning implementations, which primarily emphasize traditional quality characteristics, and the security-focused mindset held by stakeholders responsible for operational security in businesses that incorporate machine learning solutions reinforced by real-world examples of adversarial machine learning attacks [6]. This gap highlights the need for a more holistic approach to designing and deploying machine learning systems, taking into account not only their performance but also their resilience to adversarial attacks and other security challenges [7], [8].

Furthermore, we have found that the field of rough sets theory (RST) has not been thoroughly explored when it comes to its capability in attack detection. One of defining characteristics of RST is that it can be used to handle uncertainty and vagueness of data [9]. By approximating the decision boundaries of a classifier model, rough sets can be used to identify regions in the input space where adversarial perturbations are likely to occur [10]. By monitoring these regions, unusual deviations or patterns in input data can be flagged as potential adversarial attacks. This approach, if proved to be working, can not only provide a robust mechanism for detecting adversarial examples but also offer insights into the underlying structure of the data and its susceptibility to manipulations, thereby informing the design of more secure machine learning models. In this work, we want to test the usefulness of RST methods in practical security applications in the domain of adversarial machine learning prevention.

The end goal of our work is to create a robust black-box-based method that can be utilized in real-world scenarios for the detection and prevention of misclassification adversarial attacks on machine learning models, increasing the safety and trustworthiness of machine learning applications in everyday scenarios.

II. Related Work

A. Advesarial Machine Learning (AML)

Starting with the pioneering works of Szegedy et al. [11] and Goodfellow et al. [12], the topic of adversarial machine learning has entered the spotlight of the research community. Those works demonstrated that it is possible to influence the operation of machine learning models, most notably image-based classifiers, by adding limited amplitude (undetectable to the human eye) perturbances to original images, causing spectacular cases of misclassification of images.

Since the concept's inception, it has left the walls of academia, and real-world adversarial machine-learning attacks have been proven possible in various areas [13], [14].

There are several possible ways to classify the diverse world of AML attacks [3], [8], [15]. The classification of AML attacks is based on a different axis:

- Knowledge-based classification — distinguishes attacks based on the amount of knowledge an attacker has about the target model.
- Capability-Based Classification — considers the capabilities of the attacker and the stage of the machine learning pipeline targeted.

Topical area: Advanced Artificial Intelligence in Applications

- Goal-Based Classification — differentiates attacks based on the attacker's objectives.

A point of note is that most of the published papers refer to attacks and defenses on image data [15]. Only in recent years, the interest in attacks and defense on tabular data processing models has increased [16].

B. AML detection

Complementary to works dedicated to increasing the robustness of models against adversarial machine learning, significant effort is put into the detection of attacks against ML models. These techniques are primarily designed to identify inputs that have been modified with the intent of misleading a machine-learning model.

Some detection strategies attempt to detect adversarial examples by identifying instances that significantly deviate from the distribution of normal instances. An example of such a technique, specific for adversarial attacks on image classification models, has been described in [17]. The detection technique presented therein hangs on the realization that adversarial images place abnormal emphasis on the lower-ranked principal components from principal component analysis (PCA), which allows adversarial examples to stand out after PCA whitening. Most recently, salience-based methods have been used to analyze adversarial examples for NLP models — based on an observation that salience tokens have a direct correlation with adversarial perturbations [18].

Another approach to the detection of adversarial perturbations is to train a separate classifier used to classify inputs as normal or adversarial. In [19], such an approach was implemented using neural network classifiers. The method has been proven to be useful for the detection of small adversarial perturbances in images (below the human-detection threshold). This auxiliary classifier can be integrated with the main model and can provide a reasonable level of adversarial threat detection [20].

An interesting approach for the detection of perturbed images has been presented in [21], where a method has been presented that detects adversarial examples by comparing the output of a discriminator of a generative adversarial network (GAN) trained on the dataset — with the realization that adversarial examples are scored lower by the discriminator part of the GAN.

III. Notions and definitions

A. Adversarial attack types used

In this work, we have tested our attack detection method against three known attack techniques: HopSkipJump, PermuteAttack, and ZOO. These attack methods were chosen based on three criteria: a thorough description of the attack methodology in an academic paper, its applicability to attacks on classifier models operating on tabular data, and the availability of its source code. While the choice of attack methods to be used in our work was arbitrary, it was considered to be proper for the preliminary attack detection method verification presented in this work.

1) HopSkipJump Attack: The HopSkipJump attack, also known as the Decision-Based Boundary attack, is an adversarial attack on machine learning models designed to generate adversarial examples by directly manipulating the input data to cause misclassifications while minimizing the perturbation to the original input [22].

It is an iterative, decision-based attack, meaning that it only requires access to the model's output decisions (e.g., classification labels) rather than full access to the model's internal workings or gradients. The attack algorithm consists of three main steps:

- Hop: Initialization of the adversarial example by searching for a starting point near the decision boundary of the model.
- Skip: Binary search along the line connecting the original input and the initialized adversarial example to find a point that lies closer to the decision boundary.
- Jump: Gradient-free optimization to further perturb the adversarial example while keeping it within a predefined perturbation budget.

2) Permute Attack: PermuteAttack, described in [23], is a counterfactual example generation method capable of handling tabular data including discrete and categorical variables. The method is based on gradient-free optimization genetic algorithm, that permutes randomly selected features making sure that resulting values are within ranges that are not outstanding for a given data set. As a result, it produces adversarial data points that are modified, as compared to the original data points, in a way that can elude some anomaly-detecting methods. Resulting adversarial examples can be also used for the analysis of the robustness of the attacked model.

3) Zeroth-Order Optimization (ZOO) Attack: The Zeroth Order Optimization (ZOO) attack is a black-box adversarial attack proposed by Chen et al. [24] The key idea behind the ZOO attack is to approximate gradients of the target model using zeroth-order (derivative-free) optimization methods, allowing the attacker to generate adversarial examples without direct access to the model's gradients or architecture. The ZOO attack steps:

- Approximate the gradients using zeroth-order optimization, such as the coordinate-wise finite-difference method or the spherical coordinate-based method.
- Compute the adversarial perturbation using the approximated gradients.
- Apply the perturbation to the original input, ensuring that the adversarial example remains within a predefined perturbation budget.

B. Diagnostic attributes

The whole workflow connected with model approximation and diagnostic attributes was originally described in work [25]. Here we just shortly call the main idea. This approach focuses on building a surrogate model for origin model predictions using the rough sets theory [26]. Based on discretized input data set we construct the ensemble of approximate reducts. The next step is to create a neighborhood for every instance

in the diagnosed data set as a set of instances from the train data set that is similar to a given instance in the diagnosed data set. The defined neighborhood is a basis for calculating the diagnostic attributes listed below.

- Target consistency with approximations in neighborhood — measuring the consistency of the target of the diagnosed instance with the approximations from the neighborhood of this instance.
- Prediction consistency with targets in the neighborhood — measuring the consistency of the prediction of the diagnosed instance with the targets from the neighborhood of this instance.
- Target consistency with targets in neighborhood — measuring the consistency of the target of the diagnosed instance with the targets from the neighborhood of this instance.
- Targets and approximations inconsistency in neighborhood — measuring the inconsistency of targets and approximations in the neighborhood of diagnosed instance.
- Targets diversity in the neighborhood — measuring the diversity of targets in the neighborhood of diagnosed instance in comparison to the diversity of targets calculated on the whole diagnosed data set.
- Approximations diversity in the neighborhood — measuring the diversity of approximations in the neighborhood of diagnosed instance in comparison to the diversity of approximations calculated on the whole diagnosed data set.
- Uncertainty — the measure of uncertainty of prediction based on the approximations.
- Neighborhood size — the number of instances in the neighborhood of diagnosed instance.

We used the Kolmogorov-Smirnov (KS) test [27] to compare the distribution of diagnostic attributes. Additionally, the Wilcox signed rank test [28] for paired two samples was conducted. The first test compares the distance between distributions while the second measure only changes in the location parameter.

IV. Experiments and results

To evaluate proposed diagnostic attributes in attack detection we prepare benchmark data sets. From OpenML[1] we gathered 22 data sets with classification task. Each data set was split into train and diagnosed parts assuming that the diagnosed data set should consist of at least 100 observations. A list of data sets is placed in the appendix in Table IV.

For each data set, we fitted a logistic regression model, support vector machine, and XGBoost. Afterward, three adversarial attacks were conducted at the diagnosed part of each data set.

Figure 1 shows the distribution of balanced accuracy measured at the diagnosed data set for the origin (base) model and how it changed after the given attack.

[1]https://www.openml.org/

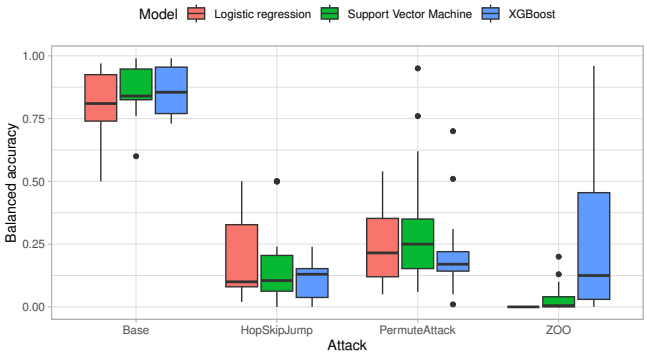

Fig. 1. Distribution of balanced accuracy in analyzed datasets across the type of model and attack

It can be seen that post-attack models in most cases result in worse performance than the base model. The median balanced accuracy for all base models is above 0.8 while in the case of the HopSkipJump attack, it is around 0.1. For Permute attack median value is slightly higher and equal to around 0.2. In the ZOO attack, these values are close to 0, but high dispersion of results for the XGBoost model can be observed.

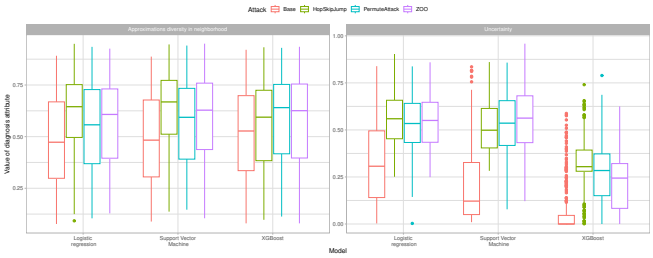

Fig. 2. Distribution of selected diagnostic attributes for the spambase data set

We calculated diagnostic attributes for each analyzed data set and attack, resulting in 264 tables with results (22 data sets $\times$ 3 model types $\times$ 4 attack variants (no attack + 3 others)). The distribution of two diagnostic attributes for the selected data set (spambase) is presented in figure 2. We used the Kolmogorov-Smirnov test to verify the null hypothesis that there is no difference between the distribution of the given diagnostic attribute before and after the attack. We also used the Wilcoxon test to examine the hypothesis that the median of differences between the paired attributes is zero. Both of these tests indicates whether there is a significant difference between diagnostic attribute before and after the attack. We summarize this data by calculating the fraction of cases in which the null hypothesis of a given statistical test was rejected at significance level $\alpha = 0.05$. Results are presented at three levels of aggregation — effectiveness of detecting attacks at the type of attack (table I), the type of model (table II), and diagnostic attribute (table III).

In the case of the HopSkipJump attack KS test rejected the null hypothesis in 88% cases while the Wilcox test in 95%.

TABLE I
PERCENTAGE OF DETECTED DIFFERENCES BETWEEN DIAGNOSTIC ATTRIBUTES AT ATTACK LEVEL

Attack type	Kolmogorov-Smirnov	Wilcox
HopSkipJump	88.28	94.91
PermuteAttack	79.36	94.63
ZOO	81.25	94.83

The Permute attack and ZOO attack were slightly harder to detect — the effectiveness of the Kolmogorov-Smirnov test is 79% and 81% respectively. For the Wilcox test, values are close to 95%.

TABLE II
PERCENTAGE OF DETECTED DIFFERENCES BETWEEN DIAGNOSTIC ATTRIBUTES AT MODEL LEVEL

Model type	Kolmogorov-Smirnov	Wilcox
Logistic regression	84.47	92.59
Support Vector Machine	85.61	94.24
XGBoost	78.52	97.75

At the model type level, we detected 79% of attacks conducted on the XGBoost model, almost 84% on logistic regression, and 86% on SVM using the Kolmogorov-Smirnov test. With Wilcox test success rate is equal to 93% for Logistic regression, 94% for SVM, and 98% for the XGBoost model.

TABLE III
PERCENTAGE OF DETECTED DIFFERENCES BETWEEN DIAGNOSTIC ATTRIBUTES AT ATTRIBUTE LEVEL

Diagnostic attribute	Kolmogorov-Smirnov	Wilcox
Approximations diversity in neighborhood	87.76	98.96
Neighborhood size	72.45	96.35
Prediction consistency with targets in neighborhood	81.12	83.85
Target consistency with approximations in neighborhood	90.31	98.44
Target consistency with targets in neighborhood	87.24	95.83
Targets and approximations inconsistency in neighborhood	60.71	89.58
Targets diversity in neighborhood	88.27	96.35
Uncertainty	95.41	98.95

Another issue was the verification of diagnosis attributes effectiveness in attack detection. According to the Kolmogorov-Smirnov test, the highest detection rate was obtained for uncertainty (95%), target consistency with approximations in the neighborhood (90%), and targets diversity in the neighborhood (88%). In the case of the Wilcox test, we obtain similar high results for three attributes: uncertainty, approximations diversity in the neighborhood, and target consistency with approximations in the neighborhood.

V. CONCLUSIONS

In this paper, we have presented a novel approach to detecting adversarial attacks on machine learning classification models operating on tabular data. By analyzing differences in diagnostic parameters calculated on an approximated representation of the model under attack, we demonstrate that adversarial attacks can be detected in a statistically significant manner. Experiments performed on real-world datasets confirm the effectiveness of our method and its potential for further development as a detection technique for adversarial attacks.

A. Limitations

The method developed and presented in this paper has several limitations, which will be tackled in future works. Most notably:

- The robustness of the method to attack variability has not been subject to wider assessment — for the initial method validation a sample of three attacks was chosen, but it does not cover the range of currently known and published attacks on models designed for processing of tabular data.
- The method assumes that the model being monitored is replicable with a rough-sets-based method presented in previous work. In this paper, it was verified on three classification models, with the assumption that the underlying model replication method provides a layer of abstraction that is strong enough to consider our method model-agnostic. Verification of this hypothesis has not been a subject of this work.
- The computational efficiency optimization and scalability have not been, by design, within the scope o the work presented herein.
- The method has not been benchmarked against available AML detection techniques.

B. Future Work

Future work will be streamlined into three distinctive work streams.

First, we will broaden the range of scenarios on which the method is tested. The method will be verified on a larger representation of known attack methods and exploration of their attack parameter space. Special attention will be given to methods that attempt low-profile adversarial attacks, attempting to pass under the detection threshold of traditional monitoring tools. We also plan to compare our approach with other methods which aim to detect AML. Furthermore, we will verify the assumption of the method being model-agnostic, by checking how its effectiveness changes when used on a different original model being attacked.

The second workstream will be devoted to new features of the method:

- Concept drift detection - examining differences in diagnostic attributes behavior between changes in data resulting from malicious attacks and different types of concept drifts - both stochastic and deterministic in nature
- Exploration of possibility for new diagnostic attributes definition. Specifically - looking for diagnostic attributes that increase specificity and sensitivity of attack detection heuristics

In the third work stream, we intend to analyze the scalability of the method and prepare a thorough comparison of the

presented attack detection method with available alternative adversarial attack detection methods.

We will also consider extending the set of diagnostic attributes with information obtained on the basis of approximation of diagnosed models with white-box models (e.g. rule-based models [29])

REFERENCES

[1] M. Kozielski, M. Sikora, and Ł. Wróbel, "Disesor-decision support system for mining industry," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2015, pp. 67–74.

[2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015. doi: 10.1126/science.aaa8415. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aaa8415

[3] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018. doi: 10.1109/ACCESS.2018.2807385

[4] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," pp. 506–519, 04 2017. doi: 10.1145/3052973.3053009

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," pp. 39–57, 05 2017. doi: 10.1109/SP.2017.49

[6] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[7] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[8] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," p. 2154–2156, 2018. doi: 10.1145/3243734.3264418. [Online]. Available: https://doi.org/10.1145/3243734.3264418

[9] Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data*. Springer Science & Business Media, 1991.

[10] A. Skowron and L. Polkowski, *Rough sets in knowledge discovery 1: Basic concepts*. CRC Press, 1998.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2013.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[13] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 07 2016.

[14] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *CoRR*, vol. abs/1707.08945, 2017. [Online]. Available: http://arxiv.org/abs/1707.08945

[15] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020. doi: https://doi.org/10.1016/j.eng.2019.12.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S209580991930503X

[16] K. Kireev, B. Kulynych, and C. Troncoso, "Adversarial robustness for tabular data through cost and utility awareness," in *NeurIPS ML Safety Workshop*, 2022. [Online]. Available: https://openreview.net/forum?id=3ieyhWF1Hk

[17] D. Hendrycks and K. Gimpel, "Early methods for detecting adversarial images," *arXiv preprint arXiv:1705.07263*, 2017.

[18] L. Li, X. Chen, Z. Bi, X. Xie, S. Deng, N. Zhang, C. Tan, M. Chen, and H. Chen, "Normal vs. adversarial: Salience-based analysis of adversarial samples for relation extraction," in *Proceedings of the 10th International Joint Conference on Knowledge Graphs*, ser. IJCKG '21. New York, NY, USA: Association for Computing Machinery, 2022. doi: 10.1145/3502223.3502237. ISBN 9781450395656 p. 115–120. [Online]. Available: https://doi.org/10.1145/3502223.3502237

[19] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "Detecting adversarial perturbations with neural networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[20] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. Mcdaniel, "On the (statistical) detection of adversarial examples," *ArXiv*, vol. abs/1702.06280, 2017.

[21] G. K. Santhanam and P. Grnarova, "Defending against adversarial attacks by leveraging an entire gan," 2018.

[22] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," *arXiv preprint arXiv:1904.02144*, 2019.

[23] M. Hashemi and A. Fathi, "Permuteattack: Counterfactual explanation of machine learning credit scorecards," 2020.

[24] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[25] A. Janusz, A. Zalewska, Łukasz Wawrowski, P. Biczyk, J. Ludziejewski, M. Sikora, and D. Ślęzak, "Brightbox—a rough set based technology for diagnosing mistakes of machine learning models," *Applied Soft Computing*, p. 110285, 2023. doi: https://doi.org/10.1016/j.asoc.2023.110285. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623003034

[26] A. Skowron and D. Ślęzak, "Rough Sets Turn 40: From Information Systems to Intelligent Systems," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems, FedCSIS 2022, Sofia, Bulgaria, September 4-7, 2022*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 30, 2022. doi: 10.15439/2022F310 pp. 23–34. [Online]. Available: https://doi.org/10.15439/2022F310

[27] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[28] R. C. Blair and J. J. Higgins, "Comparison of the power of the paired samples t test to that of wilcoxon's signed-ranks test under various population shapes." *Psychological Bulletin*, vol. 97, no. 1, p. 119, 1985.

[29] A. Gudyś, M. Sikora, and Ł. Wróbel, "Rulekit: A comprehensive suite for rule-based learning," *Knowledge-Based Systems*, vol. 194, p. 105480, 2020.

APPENDIX

TABLE IV
BASIC CHARACTERISTICS OF DATA SETS USED IN EXPERIMENTS. THE COLUMNS N, $|A|$, AND $|L|$ SHOW THE TOTAL NUMBER OF INSTANCES, ATTRIBUTES, AND CLASSES, RESPECTIVELY.

name	N	$\|A\|$	$\|L\|$
Bioresponse	3751	1776	2
churn	5000	20	2
cmc	2000	47	10
cnae-9	1080	856	9
dna	3186	180	3
har	10299	561	6
madelon	2600	500	2
mfeat-factors	2000	47	10
mfeat-fourier	2000	76	10
mfeat-karhunen	2000	47	10
mfeat-zernike	2000	47	10
nomao	34465	118	2
optdigits	2000	47	10
pendigits	10992	16	10
phoneme	5404	5	2
qsar-biodeg	1055	41	2
satimage	6430	36	6
semeion	1593	256	10
spambase	2000	47	10
wall-robot-navigation	5456	24	4
wdbc	569	30	2
wilt	4839	5	2