COMPUTER SCIENCES AND INFORMATION TECHNOLOGIES

UDC 004.855.5(045)
DOI:10.18372/1990-5548.71.16816

[1]V. M. Sineglazov,
[2]K. S. Lesohorskyi

ON NOISE EFFECT IN SEMI-SUPERVISED LEARNING

[1]Aviation Computer-Integrated Complexes Department, Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine
[2]Faculty of Informatics and Computer Science, National Technical University of Ukraine "Ihor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
E-mails: [1]svm@nau.edu.ua ORCID 0000-0002-3297-9060, [2]lesogor.kirill@gmail.com

***Abstract*—*The article deals with the problem of noise effect on semi-supervised learning. The goal of this article is to analyze the impact of noise on the accuracy of binary classification models created using three semi-supervised learning algorithms, namely Simple Recycled Selection, Incrementally Reinforced Selection, and Hybrid Algorithm, using Support Vector Machines to build a base classifier. Different algorithms to compute similarity matrices, namely Radial Bias Function, Cosine Similarity, and K-Nearest Neighbours were analyzed to understand their effect on model accuracy. For benchmarking purposes, datasets from the UCI repository were used. To test the noise effect, different amounts of artificially generated randomly-labeled samples were introduced into the dataset using three strategies (labeled, unlabeled, and mixed) and compared to the baseline classifier trained with the original dataset and the classifier trained on the reduced-size original dataset. The results show that the introduction of random noise into the labeled samples decreases classifier accuracy, while a moderate amount of noise in unmarked samples can have a positive effect on classifier accuracy.*

Index Terms—Data noise; machine learning; semi-supervised learning; support vector machines.

I. INTRODUCTION

Semi-supervised learning is a machine learning approach that leverages both labeled and unlabeled data for model training. Usually, a dataset is split unevenly with unlabeled data being the majority of samples. This leads to the problem of selecting helpful unlabeled samples that can increase the model accuracy.

To analyze the effect of noise on the semi-supervised learning algorithm it is best to look at well studied problem of binary classification. Existing meta-semi-supervised algorithms such as Simple Recycled Selection (SRS) [1], incrementally Reinforced Selection (IRS) [2], and Hybrid Algorithm (HYB) introduced in [3] as they leverage a supervised learning algorithm to train the model, which enables the comparison of baseline supervised algorithm with the semi-supervised algorithm. These algorithms assign pseudo labels based on similarity kernel and select samples with the most bias towards one of the classes and use these samples to expand the training set. These algorithms are based on the assumptions of clustering and smoothness. Smoothness assumption assumes that points that are located close in the dataspace are more likely to be in the same class, while cluster assumption assumes that points that belong to one class are likely to form a group or cluster [9]. This assumptions enable pseudo-lable assignment to the unlabeled data by measuring distance to the nearby labeled data points and selecting high-confidence "strong" unlabeled points to be the part of the closest labeled cluster.

However, both labeled and unlabeled data can contain noise, which can increase or decrease the model accuracy.

This paper aims to analyze the effect of noise on the met-semi-supervised learning algorithms, namely SRS, IRS, and HYB, and compare it to the impact on the baseline Support Vector Machine (SVM) [4].

II. PROBLEM STATEMENT

The problem of semi-supervised learning is the problem of training classifier H leveraging labeled dataset $L = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$ and unlabeled dataset $U = \{x_1, x_2, \ldots x_m\}$.

This paper aims to research the effect of noisy data on semi-supervised learning algorithms by introducing artificially generated noise into labeled datasets and dropping a certain percentage of the labels.

The noise model we use to augment source datasets creates new noisy samples by randomly combining features of existing samples together and assigning a random label to the sample. The details of the noise algorithm are expanded in section 4.

Our goal is to evaluate the performance of supervised and semi-supervised algorithms by measuring the classifier accuracy after training it on datasets with different levels of noise.

To evaluate the effect of the noise on model quality, we measure the error rate of the presented algorithms for binary classification problems and compare it to the baseline SVM classifier. In this paper we use the binary classification error rate formula:

$$e = 1 - \frac{P_t + N_t}{P_t + N_t + P_f + N_f}, \quad (1)$$

where P_t is a quantity of true positives; N_t is a quantity of true negatives; P_f is a quantity of false positives; N_f is a quantity of false negatives.

When introducing noise into the dataset for semi-supervised algorithms, three strategies are used: introduce the noise into the labeled data only (labeled noise), introduce the noise into the unlabeled data only (unlabeled noise), and introduce the noise into both labeled and unlabeled data (mixed noise).

The noise agent generates new noisy samples by combining features of randomly selected existing samples and assigning random labels:

$$\text{noise}_{ij} = \text{random}(X)_j,$$
$$\text{labels}_i = \text{random}\left(\{-1,1\}\right),$$

where X is the original samples; X_j is the feature j of X.

Generated samples are then introduced into datasets based on the noise introduction strategy (labeled, unlabeled, mixed) and classifier is trained using one of the SSL algorithms.

To study the effect of noise semi-supervised algorithms were selected because they enable the usage of a supervised model which can be used to establish the baseline accuracy. Among the semi-supervised algorithms SRS, IRS, and HYB algorithms enable the usage of a base supervised model by assigning pseudo labels to the unlabeled data and leveraging the supervised learning algorithm to create a classifier. To assign the pseudo labels to the unlabeled data, a similarity matrix is built using a kernel function. RBF and cosine similarity kernel functions were chosen because of their wide adoption, while KNN was chosen as it promises an improvement of the underlying kernel function.

Support Vector Machine was chosen as a baseline supervised classifier as it is well suited and well suited for binary classification problems.

III. Related Works

A. Simple Recycled Selection

Simple Recycled Selection algorithm introduced in [1] is based on the idea of iteratively improving the model accuracy by combining labeled data L with small subset U_t of unlabeled data U.

To select the strong unlabeled samples similarity matrix S and predicted labels H are used to compute the confidence level of value being assigned to positive class p or negative class q. Similarity matrix is a pairwise matrix of labeled and unlabeled samples with values in range of 0 to 1 inclusive, where 1 means exact match and 0 – lack of any similarity between the two samples.

z and z_{conf} is then computed as $z = \text{sign}(p - q)$ and $z_{\text{conf}} = |p - q|$. z will be used as a pseudo label – if $z_0 = -1$, then U_i is considered to be of a negative class and vise versa. z_{conf} is a measure of how "strong" unlabeled sample is.

Unlabeled data U is sorted by the confidence level z_{conf} in the descending order and top k samples are selected with their respective pseudo labels and combined into the U_t.

The model M_i is trained using training set $P_t = U_t \cup L$. Afterward, the model is used to assign predicted labels H for U.

If the ensemble version of algorithm is used model weight w_i is calculated and prediction function $H(X)$ is updated as $H(X) = H(X) + w_i \cdot M_i(X)$.

It is important to note, that SRS relies on iteratively using a small number of unlabeled samples iteratively improving the accuracy. This means that the training set size is constant throughout the training process, so the algorithm benefits not from the unlabeled set size but rather from a few strong samples. This algorithm is better suited for problems with either relatively small unlabeled data set or if unlabeled dataset contains few strong distinct samples.

B. Incrementally Reinforced Selection

Incrementally Reinforced Selection [2] is similar to SRS in many ways. The only major difference is that, unlike SRS, IRS expands the labeled data set with selected strong unlabeled data U_t. This enables IRS to benefit from bigger unlabeled data set size, not just from few strong samples.

Since IRS will expand training set, it is well suited for problems with the vast unlabeled dataset as it will benefit from the expanded data set. However, caution should be taken when selecting the training hyper parameters as after selecting strong samples algorithm will pick weak samples which could decrease the model accuracy.

C. Hybrid Algorithm

Hybrid algorithms attempt to combine principles of both SRS and IRS together to create a more robust algorithm that picks the strongest possible unlabeled data U_t on each iteration while also expanding the the number of semples selected from the unlabeled data set.

Hybrid algorithm introduced in [3] is more universal than SRS and IRS as it can both leverage few strong samples to improve initial accuracy and then use more available unlabeled samples as it increases the unlabeled ratio. It is applicable in scenarios when both IRS and SRS are used and can prove to be more efficient, however, it requires more fine tuning for both SRS and IRS phases.

D. Kernel Functions

All of the reviewed algorithms use a similarity matrix to assign pseudo-labels to the unlabeled dataset. In our study, we used three kernel functions to calculate similarity matricies to test the impact of noise on the similarity kernels.

We experimented with three similarity kernels - Radial Bias Function [5], Cosine Simmilarity [6] and K-Nearest Neighbours [7].

Radial Bias Function (RBF) – is a function that calculates the similarity of the two samples based on squared Euclidian distance adjusted by a free parameter coefficient.

$$K(x,x') = \exp\left(\frac{\| x - x' \|^2}{2\sigma^2}\right),$$

where σ is a free parameter

Cosine Similarity is also a common choice. It is based on the cosine of the angle between two vectors created between the two samples and is calculated as

$$K(x,x') = \frac{xx'^T}{\| x \| \| y \|}.$$

K-Nearest Neighbors kernel can be considered a meta-kernel as it utilizes other kernels to improve their accuracy.

$$K'(x,x') = K(x,x) - 2K(x,x') + K(x',x'),$$

where K is an underlying similarity kernel.

In our study, we used RBF as the underlying kernel for K-Nearest Neighbors.

Usually, similarity kernels are selected based on the geometry of the problem. RBF is a reasonable and common default choice. Also, an automated hyperparameter search can be used to determine the best similarity kernel and kernel's parameters as described in [8].

IV. Experiment Setup

To study the effect of noise we designed the dataset processing method and noise generating method. All of the algorithms used in the paper are designed for the problems of binary classification, so all of the datasets are reduced to two-class datasets by selecting two majority classes.

We use labeled datasets as a baseline. To generate a semi-supervised dataset we randomly select data points in the original dataset and remove labels. In our setup, we keep 30% of points labeled while removing labels for the other 70% of the points.

During the testing, we conducted several runs with different noise levels as introduced in formula (2) and evaluated the results using the error rate formula (1). We experiment with 0, 5, 10, 20, and 30 percent of noise samples in datasets for baseline SVM, labeled data, unlabeled data, and mixed data.

Additionally, we experiment with several ways of introducing noise into the dataset as illustrated in Fig. 1.

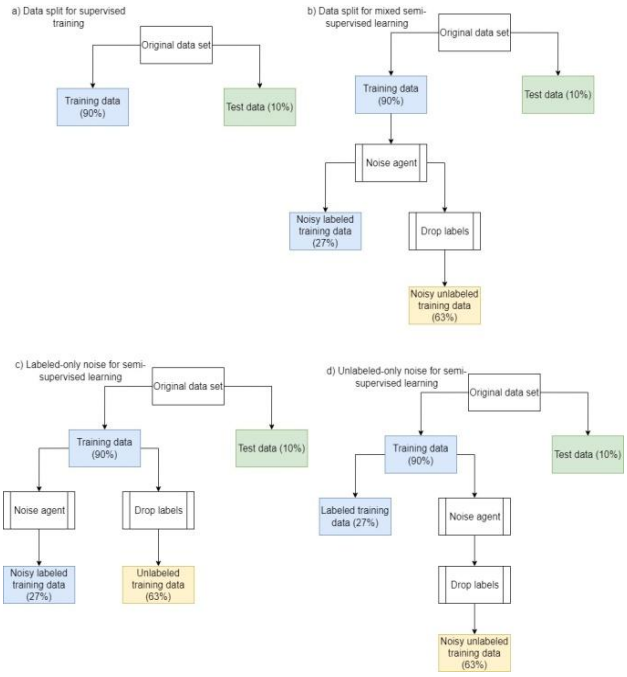

Fig. 1. Data split for different experimentation setups

Baseline experiments were conducted in the mixed mode (Fig. 1b). In this mode, we introduce noise both to the labeled and unlabeled samples. To assess the impact of noise type we additionally experimented with labeled-only dataset noise (Fig. 1c) and unlabeled-only noise (Fig. 1d) and compared the resulting error rates.

It's worth noting that we used relatively small academic datasets with clear decision boundaries, however, in real-world datasets it is important to make sure that datasets meet the assumptions of selected SSL algorithms (smoothness and clustering) in our case. Another important property that the dataset should satisfy is that the density probability function should match for both labeled and unlabeled datasets, otherwise model accuracy can decrease as SSL algorithms are likely to mislabel the unlabeled samples. Finally, as we used SVM as our baseline supervised classifier it is important that datasets have high and low-density regions in order to find a proper decision boundary. If the dataset has only high or low-density regions or class overlaps SVM will show poor performance as it would struggle to find the appropriate decision boundary.

V. Results

We tested 3 meta-semi-supervised learning algorithms and 1 supervised learning algorithm on several UCI [10] datasets. The datasets used are illustrated in Table I.

Table I. Used Datasets

Dataset	Number of attributes	Number of instances	Number of classes
balance	4	625	3
glass	10	214	7
heart	14	270	2
iris	4	150	3

To mitigate randomness in the data augmentation process each test was run 20 times and error rates were averaged afterward. Each semi-supervised test was performed with 3 different kernels – cosine, radial bias function and *k*-nearest-neighbors.

Each dataset was reduced to two dominant classes as our implementations of the algorithms are only suitable for binary classification. The results of the experiments are given in Tables II to V. The tables contain error rate values for specified percentage of noisy samples using mixed noise algorithm (Fig. 1b).

Table II. Results for the "Balance" Dataset

Algoritm / Noise Level	Base	5%	10%	20%	30%
SVM	0.0172	0.0492	0.0156	0.1429	0.1333
SVM (30% data)	0.0172	–	–	–	–
SRS (COS)	0.1575	0.1929	0.1885	0.2495	0.2376
SRS (KNN)	0.0799	0.0819	0.1297	0.1457	0.1956
SRS (RBF)	0.0569	0.061	0.1052	0.1875	0.2044
IRS (COS)	0.3937	0.4033	0.3823	0.4563	0.4429
IRS (KNN)	0.1816	0.1973	0.2344	0.2486	0.2712
IRS (RBF)	0.1023	0.1374	0.226	0.2659	0.3659
HYB (COS)	0.3144	0.3132	0.3099	0.3841	0.3403
HYB (KNN)	0.0713	0.0846	0.1099	0.176	0.2013
HYB (RBF)	0.2345	0.2341	0.2562	0.2913	0.3628

Table III. Results for "Glass" Dataset

Algoritm / Noise Level	Base	5%	10%	20%	30%
SVM	0	0.125	0.1176	0.2222	0.2253
SVM (30% data)	0	-	-	-	-
SRS(COS)	0.1682	0.1125	0.226	0.3093	0.2948
SRS(KNN)	0.0227	0.0729	0.092	0.1204	0.1879
SRS(RBF)	0.1477	0.125	0.284	0.2352	0.2052
IRS(COS)	0.3023	0.4458	0.314	0.3352	0.4052
IRS(KNN)	0.0682	0.1875	0.21	0.2611	0.3086
IRS(RBF)	0.3	0.4542	0.326	0.3519	0.3948
HYB(COS)	0.2886	0.2771	0.286	0.363	0.3948
HYB(KNN)	0.0341	0.0146	0.008	0.1111	0.1152
HYB(RBF)	0.2568	0.25	0.378	0.3278	0.4345

Table IV. Results for "Heart" Dataset

Algoritm / Noise Level	Baseline	5%	10%	20%	30%
SVM	0.2963	0.3448	0.2333	0.4545	0.4167
SVM (30% data)	0.2963	-	-	-	-
SRS (COS)	0.4305	0.4535	0.42	0.4061	0.4764
SRS (KNN)	0.3793	0.5023	0.3678	0.4714	0.4708
SRS (RBF)	0.3988	0.4267	0.3578	0.4551	0.4349
IRS (COS)	0.472	0.4919	0.48	0.4694	0.4528
IRS (KNN)	0.4427	0.4384	0.4244	0.4408	0.4528
IRS (RBF)	0.3646	0.5186	0.4078	0.401	0.3811
HYB (COS)	0.5061	0.4989	0.4898	0.4826	0.4953
HYB (KNN)	0.3817	0.4233	0.4233	0.4245	0.4915
HYB (RBF)	0.361	0.3826	0.4022	0.3378	0.45

Table V. Results for "Iris" Dataset

Algoritm / Noise Level	Baseline	5%	10%	20%	30%
SVM	0	0	0.0909	0.0833	0.0769
SVM (30% data)	0	-	-	-	-
SRS(COS)	0.0333	0.0125	0.0441	0.0111	0.015
SRS(KNN)	0	0.0125	0.0029	0.1111	0.13
SRS(RBF)	0.02	0.0	0.1853	0.1972	0.18
IRS(COS)	0.2233	0.1688	0.0912	0.2	0.24
IRS(KNN)	0	0.0156	0.0912	0.1667	0.085
IRS(RBF)	0.2333	0.2562	0.0912	0.1722	0.255
HYB(COS)	0.28	0.2969	0.3824	0.6028	0.5275
HYB(KNN)	0	0	0.0588	0.1111	0.1075
HYB(RBF)	0.41	0.375	0.6029	0.6528	0.4775

From the results several observations can be highlighted.

- KNN similarity kernel shows best performance on high dimension data, while RBF shows comparable or better performance for low dimension datasets.
- SRS seems to be the best SSL method, followed by HYB and IRS.
- Small amounts of noise (5–10%) can improve model performance. Increasing noise levels seem to have a decaying effect on the model accuracy.
- SSL methods have a high variance in error rates with noise in the data. It is caused by noise split between labeled and unlabeled data. When noise is added into unlabeled data it has a high chance of being dropped if it has low confidence values or used if it can is considered a strong example. However when noise is added to labeled data will skew confidence values for unlabeled data, introducing bias.
- Overall, on analyzed datasets it can be seen that SSL has little to no improvement compared to the original model. Small amounts of noise seem to have positive effects on high-dimension high-volume datasets, such as the heart.

Additionally, tests with different noise introduction algorithms were performed on the heart dataset to test the impact of different noise types. Several algorithms, namely SRS, boosted variation of SRS, IRS, and HYB algorithms were compared using the KNN kernel from previous experiments.

The error rate for mixed data noise labeled data noise, and unlabeled data noise is presented in Figs 2, 3, and 4 respectively.

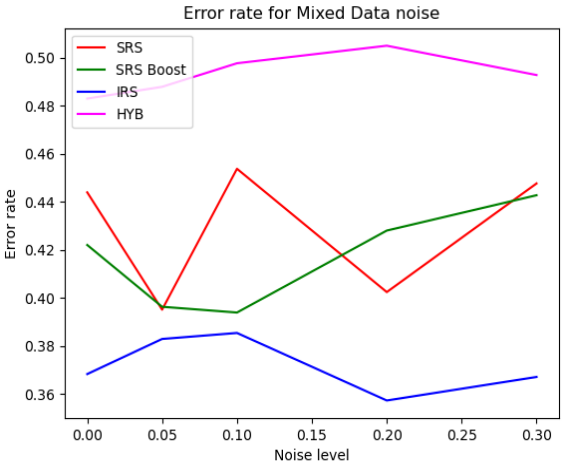

Fig. 2. Mixed data noise error rate

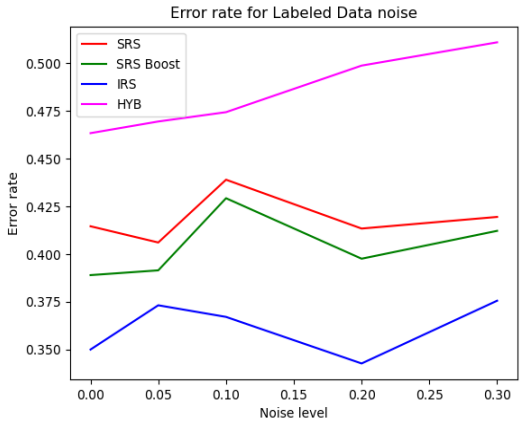

Fig. 3. Labeled data noise error rate

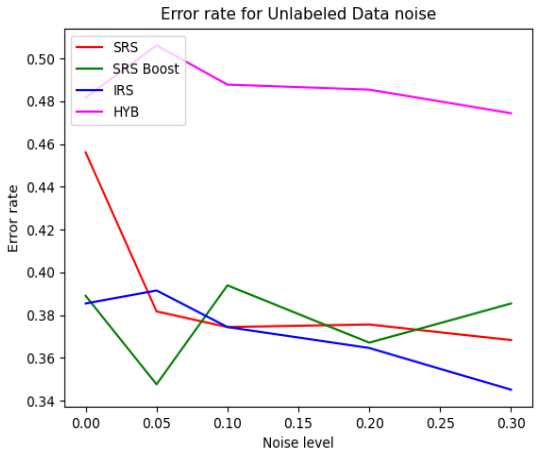

Fig. 4. Unlabeled data noise error rate

Based on this several approaches to improving the classification accuracy of semi-supervised models in the presence of noisy data. Since unlabeled data has a limited effect on classification accuracy, it is important to make sure that labeled data has as little noise as possible. A high level of noise in labeled and data algorithms that leverage few strong samples, such as SRS, show limited performance degradation. However, algorithms that leverage more unlabeled samples are more resilient to mixed noise. K-Nearest Neighbors with the underlying RBF display higher resilience to noise.

VI. Conclusions

In this paper we provided a brief overview of SRS, IRS, and HYB semisupervised-learning algorithms, an overview of KNN, RBF, and cosine similarity kernel functions, and studied the effect of noise on the models created with these algorithms.

Overall, SRS performs best for low-dimensionality datasets with a limited amount of samples, shows better accuracy, and is less sensitive to noise than other algorithms. However, on larger

datasets, it shows worse accuracy and is more susceptible to noise.

However, for high-dimensionality higher volume datasets, IRS performs better, however, it is more sensitive, especially to the mixed and labeled noise. On the lower-volume datasets, IRS performance is worse than SRS.

The hybrid algorithm shows average performance on low-noise data sets and worst performance on high volume high dimensionality datasets. Noise impact is limited, however, it also doesn't gain much accuracy from mixing noise into the unlabeled data like other methods.

Among the tested similarity kernels, KNN consistently outperforms cosine similarity and RBF and is least sensitive to noise. Cosine similarity is the most inconsistent in the noisy environment and RBF shows average results.

REFERENCES

[1] P. K. Mallapragada, et al., "SemiBoost: Boosting for semi-supervised learning," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 312, no. 11, pp. 2000–2014, Nov. 2009. https://doi.org/10.1109/TPAMI.2008.235

[2] T.-B. Le and S.-W. Kim, "On incrementally using a small portion of strong unlabeled data for semi-supervised learning algorithms," *Pattern Recognition Letters*, vol. 41, pp. 53–64, May 2014. https://doi.org/10.1016/j.patrec.2013.08.026

[3] Thanh-Binh Le, Sang-Woon Kim, "A Hybrid Selection Method of Helpful Unlabeled Data Applicable for Semi-Supervised Learning Algorithm," *IEIE Transactions on Smart Processing & Computing*, 3(4), 2014, pp. 234–239. https://doi.org/10.5573/IEIESPC.2014.3.4.234

[4] S. Suthaharan, "Support Vector Machine," *In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems*, vol. 36, pp. 207–235, 2016. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7641-3_9

[5] Orr, Mark JL, *Introduction to radial basis function networks*, 1996.

[6] Rahutomo, Faisal, Teruaki Kitasuka, and Masayoshi Aritsugi, "Semantic cosine similarity," *the 7th International Student Conference on Advanced Science and Technology (ICAST)*, vol. 4, No. 1, 2012.

[7] Yu, K., Ji, L. & Zhang, X. Kernel, "Nearest-Neighbor Algorithm," *Neural Processing Letters* 15, 147–156, 2002. https://doi.org/10.1023/A:1015244902967

[8] G. C. Cawley and N. L. C. Talbot, "Preventing over-fitting in model selection via Bayesian regularisation of the hyper-parameters," *Journal of Machine Learning Research*, vol. 8, pp. 841–861, April 2007.

[9] O. Chapelle, & A. Zien, "Semi-Supervised Classification by Low Density Separation," *In Tenth International Workshop on Artificial Intelligence and Statistics (AISTAT 2005)*, (2005). https://doi.org/10.7551/mitpress/9780262033589.001.0001

[10] D. Dua, and C. Graff, *UCI Machine Learning Repository*, 2019. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Received December 02, 2021

Sineglazov Victor. ORCID 0000-0002-3297-9060. Doctor of Engineering Science. Professor. Head of the Department.
Aviation Computer-Integrated Complexes Department, Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine.
Education: Kyiv Polytechnic Institute, Kyiv, Ukraine, (1973).
Research area: Air Navigation, Air Traffic Control, Identification of Complex Systems, Wind/Solar power plant, artificial intelligence.
Publications: more than 670 papers.
E-mail: svm@nau.edu.ua

Lesohorskyi Kyrylo. Masters Student.
Department of Information Systems, Faculty of Informatics and Computer Science, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.
Education: National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, (2022).
Research interests: artificial neural networks, artificial intelligence, distributed computing.
Publications: 1.
E-mail: lesogor.kirill@gmail.com

В. М. Синєглазов К. С. Лесогорський. Дослідження впливу шуму в напівкерованому навчанні
У статті розглядається проблема впливу шуму на точність у задачах напівкерованого навчання. Метою цієї статті є аналіз впливу шуму на точність моделей бінарної класифікації, створених за допомогою трьох напівкерованих алгоритмів навчання, а саме: Simply Recycled Selection (SRS), Incrementally Reinforced Selection (IRS) і Hybrid Algorithm (HYB). У якості базового класифікатора використано Support Vector Machine (SVM). Ми проаналізуємо різні алгоритми для обчислення матриць подібності, а саме Radial Bias Function, Cosine Simmilarity і K-Nearest Neighbours. Для цілей порівняльного аналізу використовуватимуться набори даних зі

сховища UCI. Щоб перевірити вплив шуму, різна кількість штучно згенерованих випадково позначених зразків було введено в набір даних з використанням трьох стратегій (маркована, не маркована та змішана) і порівняно з базовим класифікатором, навченим з вихідним набором даних, і класифікатором, навченим на вихідному наборі даних зменшеного розміру. Результати показують, що введення випадкового шуму в марковані зразки погіршує точність моделі, а введення випадкового шуму в немарковані дані може навпаки підвищити точність моделі.
Ключові слова: зашумлені данні; машинне навчання; напівкероване навчання; опорні векторні машини.

Синєглазов Віктор Михайлович. ORCID 0000-0002-3297-9060.
Доктор технічних наук. Професор. Завідувач кафедрою.
Кафедра авіаційних комп'ютерно-інтегрованих комплексів, Факультет аеронавігації, електроніки і телекомунікацій, Національний авіаційний університет, Київ, Україна.
Освіта: Київський політехнічний інститут, Київ, Україна, (1973).
Напрям наукової діяльності: аеронавігація, управління повітряним рухом, ідентифікація складних систем, вітроенергетичні установки, штучний інтелект.
Кількість публікацій: більше 670 наукових робіт.
E-mail: svm@nau.edu.ua

Лесогорський Кирило Сергійович. Магістр.
Кафедра інформаційних систем, Факультет інформатики та обчислювальної техніки, Національний технічний університет Україні «Київський Політехнічний Інститут імені Ігоря Сікорського», Київ, Україна.
Освіта: Національний технічний університет Україні «Київський Політехнічний Інститут імені Ігоря Сікорського», (2022).
Напрям наукової діяльності: штучні нейронні мережі, штучний інтелект, розподіленні обчислення.
Кількість публікацій: 1.
E-mail: lesogor.kirill@gmail.com

В. М. Синеглазов, К. С. Лесогорский. Исследование влияния шума в полууправляемом обучении

В статье рассматривается проблема влияния шума на точность в задачах полууправляемого обучения. Целью этой статьи является анализ влияния шума на точность моделей, созданных с помощью трех полууправляемых алгоритмов обучения, а именно: Simply Recycled Selection (SRS), Incrementally Reinforced Selection (IRS) и Hybrid Algorithm (HYB). В качестве базового классификатора использован Support Vector Machine (SVM). Проанализированы различные алгоритмы для вычисления матриц подобия, а именно Radial Bias Function, Cosine Simmilarity и K-Nearest Neighbours. Для целей сравнительного анализа будут использоваться наборы данных из хранилища UCI. Чтобы проверить влияние шумов, различные количества искусственно сгенерированных случайно размеченных образцов были введены в набор данных с использованием трех стратегий (размеченные, неразмеченные, смешанные) и сравнены с базовым классификатором, обученным исходным набором данных, и классификатором, обученным на исходном наборе данных уменьшенного размера. Результаты показывают, что ввод случайного шума в маркированные образцы ухудшает точность модели, а ввод случайного шума в немаркированные данные может наоборот повысить точность модели.
Ключевые слова: зашумленные данные; машинное обучение; полууправляемое обучение; опорные векторные машины.

Синеглазов Виктор Михайлович. ORCID 0000-0002-3297-9060.
15, Национальный авиационный университет, Киев, Украина.
Образование: Киевский политехнический институт, Киев, Украина, (1973).
Направление научной деятельности: аэронавигация, управление воздушным движением, идентификация сложных систем, ветроэнергетические установки, искусственный интеллект.
Количество публикаций: более 670 научных работ.
E-mail: svm@nau.edu.ua

Лесогорский Кирилл Сергеевич. Магистр.
Кафедра информационных систем, Факультет информатики и вычислительной техники, Национальный технический университет Украины «Киевский Политехнический Институт имени Игоря Сикорского», Киев, Украина.
Образование: Национальный технический университет Украины «Киевский Политехнический Институт имени Игоря Сикорского», (2022)
Направление научной деятельности: искусственные нейронные сети, искусственный интеллект, распределенные вычисления.
Количество публикаций: 1.
E-mail: lesogor.kirill@gmail.com