УДК: 00.005
DOI: 10.25559/SITITO.15.201902.290-297

The Effect of the ADASYN Method on Widespread Metrics of Machine Learning Efficiency

M. A. Baimakhanbetov, K. S. Nurumov, U. A. Ospanova, T. K. Buldybayev, I. G. Akoyeva*
Information-Analytical Center, JSC, Nur-Sultan, Kazakhstan
18 Dostyk Ave., Nur-Sultan 010000, Kazakhstan
* inessa.akoyeva@iac.kz

Abstract

The article presents the results of experimental work comparing the performance metrics of machine learning algorithms on imbalanced text corpora using the method of synthetic data generation ADASYN and without it. The work was carried out on an imbalanced corpus, consisting of 5,211 news texts, formed by cluster sampling for one year. The corpus annotation is produced according to the indicators of the tonality of texts by categories: neutral, positive, and negative, with a significant predominance of articles of neutral tone.

There are many widely used methods to overcome the problem of data imbalance. Often, when working with imbalanced data, the resulting accuracy provides acceptable results, but other performance indicators are low. Such contradictory results usually occur when it comes to in-depth analysis of the text in the study of social or medical phenomena. This paper shows how performance metrics of the same machine learning algorithms change when using the ADASYN method while analyzing an imbalanced text corpus using the K-nearest neighbors' method and Naive Bayes.

The study considers the issue of the application of the method and its results in solving the problem of text classification. Comparative characteristics of the machine algorithms operation before and after the application of ADASYN provide a researcher with a better understanding of which machine learning performance metrics are more suitable when working with imbalanced data.

As a result, authors present the observations and conclusions about the features of the method and put forward a number of proposals for further research in this area to compare the results obtained with the effects of the application of another method.

Keywords: ADASYN, data imbalance, machine learning, performance metrics, accuracy.

Funding: This work was supported by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan under the Grant number BR05236839.

For citation: Baimakhanbetov M.A., Nurumov K.S., Ospanova U.A., Buldybayev T.K., Akoyeva I.G. The Effect of the ADASYN Method on Widespread Metrics of Machine Learning Efficiency. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2019; 15(2):290-297. DOI: 10.25559/SITITO.15.201902.290-297

Влияние метода ADASYN на широко распространенные метрики эффективности машинного обучения

М. А. Баймаханбетов, К. С. Нурумов, У. А. Оспанова, Т. К. Булдыбаев, И. Г. Акоева*
АО «Информационно-аналитический центр» МОН РК, г. Нур-Султан, Казахстан
010000, Казахстан, г. Нур-Султан, пр. Достык, д. 18
*inessa.akoyeva@iac.kz

Аннотация

В статье представлены результаты экспериментальной работы по сравнению показателей эффективности алгоритмов машинного обучения на несбалансированных текстовых корпусах при помощи метода синтетического генерирования данных ADASYN и без него. Работа проведена на несбалансированном корпусе, состоящем из 5211 новостных текстов, сформированных методом кластерной выборки за один год. Разметка корпуса произведена по показателям тональности текстов по категориям: нейтральная, положительная и негативная, при значительном преобладании статей нейтральной тональности.

Существует множество широко используемых методов для преодоления проблемы дисбаланса данных. Часто, при работе с несбалансированными данными итоговая точность дает приемлемые результаты, но другие показатели эффективности очень низки. Такие противоречивые результаты часто имеют место, когда речь идет о глубоком анализе текста при исследовании социальных или медицинских явлений. В данной работе на примере анализа несбалансированного текстового корпуса методами К-ближайших соседей и Наивный Байес показано, как меняются распространенные метрики эффективности тех же самых алгоритмов машинного обучения при применении метода ADASYN.

Рассмотрен вопрос применения метода и его результаты при решении задачи классификации текстов. Сравнительные характеристики работы машинных алгоритмов до и после применения ADASYN предоставляют исследователю лучшее понимание того, на какие метрики эффективности машинного обучения лучше ориентироваться при работе с несбалансированными данными.

По итогам исследования представлены наблюдения и выводы авторов об особенностях применения метода и выдвинут ряд предложений по дальнейшим исследованиям в этой области для сравнения полученных результатов с эффектами от применения других методов.

Ключевые слова: ADASYN, несбалансированные данные, машинное обучение, показатели эффективности, точность.

Финансирование: данная статья подготовлена в рамках реализации ПЦФ №BR05236839 Комитета науки Министерства образования и науки Республики Казахстан.

Для цитирования: *Баймаханбетов М. А., Нурумов К. С., Оспанова У. А., Булдыбаев Т. К., Акоева И. Г.* Влияние метода ADASYN на широко распространенные метрики эффективности машинного обучения // Современные информационные технологии и ИТ-образование. 2019. Т. 15, № 2. С. 290-297. DOI: 10.25559/SITITO.15.201902.290-297

Introduction

When conducting research involving processing texts with machine learning (ML) algorithms have become a common thing among the scientific community due to the large amounts of information.
Researchers whose focus area implies a classification of texts to assess the performance of their machine learning algorithms often turns to accuracy metrics to report on the quality of their performance as to the most pervasive one. However, there is a variety of established metrics designated to evaluate the performance of machine learning such as confusion matrix [1], precision-recall [2], F1 score [3], and others that are also a handy metric to understand the ML performance quality.
Another common thing happening when conducting research is imbalanced data. When dealing with imbalanced datasets, the accuracy metrics fall short of classifying the small class items due to unequal data distribution. To address the problem, there is a set of widely used techniques to overcome the data imbalance problem [4]. In the case of imbalanced data and skewed distributions, we often face the case when the calculation of accuracy gives us acceptable results but other performance metrics are very low. Such contradictory results are often the case when it comes to text mining on extended textual information, or when researching social or medical phenomena.
This article shows how application of ADASYN method may help to understand which performance metric a researcher should rely on and report when researching imbalanced textual corpora by comparing the performance of K-Nearest Neighbors [5] and Naïve Bayes [6] algorithms before and after ADASYN [7].

Methods and data

The dataset under analysis was a collection of 5211 news articles from 5 news websites: tengrinews.kz, camonitor.kz, azattyq.org, kz.expert, zakon.kz differing from each other by topic focus, style, and level of popularity published within one year selected by systematic clustered sampling. The corpus was annotated according to the tone of the article by the coders acting under a unified instruction what to consider "neutral", "positive", and "negative" texts and articles in each corpus were classified into these three categories.
The class labeling of the corpus is shown below in Table 1.

Table 1. **Corpus labeling**

Corpus labeling	
negative	696
neutral	4120
positive	395
N total=	5211

As we can see from the above the data distribution is highly imbalanced: the positive tone articles comprise only 7.6%, and negative tone articles, 13.4% of the dataset. Articles classified as neutral constitute the largest part of the dataset amounting to 79%.
Methods applied in our research are described below.
Text mining and machine learning

Text mining methods represent automatic processing of text information which can be considered as 'digitalization' of texts [8].
To extract features from the pre-processed texts, we applied the IDF approach assessing the importance of a word for a text compared to all other documents in the corpus. The essence of it is that if a word is frequently used in one text and is not a stop-word, but at the same time is rarely used in other texts of the corpus, then its importance for this document is greater [9]. The formalized function for the approach is as follows:

$$IDF(i,j)=\begin{cases}0 & if\ wf_{i,j}=0\\ \left(1+\left(\log wf_{i,j}\right)\right)*\log\dfrac{N}{df_i} & if\ wf_{i,j}\geq 1\end{cases} \quad (1),$$

where N is the total number of documents, df_i is the frequency of documents for the i'th word[1].
The main TM algorithm finds and indexes the words in texts and then builds a frequency matrix of words use cases in each document. These frequencies are subsequently transformed into other values that indicate the relative 'significance' of the words and their semantics in initial texts based on singular value decomposition (SVD). Singular value decomposition is a method of decomposing a matrix into three other matrices:

$$A=U*S*V^T \quad (2),$$

where:
A is m×n matrix
U is m×n orthogonal matrix
S is n×m diagonal matrix
V is n×m orthogonal matrix

The use of SVD to allocate a common space of variables and observations is closely related to various statistical methods such as correspondence and factor analysis. In general, the task of this method is to reduce the overall dimension of the original matrix (the number of source documents x the number of selected words y), where each corresponding dimension represents the greatest degree of variability. In the Statistica[2] software, we used for the processing these values are assigned to the texts as concepts. This allows defining latent semantic space that organizes words and texts through determining several most important dimensions (concepts) that have the greatest variability between words and documents. Then we should define the number of those concepts should typically be small [10, 11].
K-Nearest Neighbors(K-NN) algorithm
K-Nearest Neighbors is a non-parametric ML classification method implying the input of k-closest training items in the feature space and were an item the classification is made based on similarity principle – the item is assigned to the class most common among its k-nearest neighbors[3].
The similarity is defined according to a distance metric between two data points based on Euclidean distance:

$$d(x,x')=\sqrt{(x_1-x'_1)^2+(x_2-x'_2)^2+\ldots(x_n-x'_n)^2} \quad (3)$$

[1] StatSoft. Electronic Statistics Textbook [Electronic recourse]. Available at: http://statsoft.ru/home/textbook/default.htm (accessed 10.04.2019). (In Russ.)

[2] STATISTICA [Electronic recourse]. Available at: http://statsoft.ru/ (accessed 10.04.2019). (In Russ.)

[3] Italo J. KNN (K-Nearest Neighbors) #1. How it works? Nov 8, 2018. [Electronic recourse]. Available at: https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d (accessed 10.04.2019). (In Eng.)

More formally, given a positive integer *K*, an unseen observation *x* and a similarity metric d, KNN classifier computes d between *x* and each training observation. If given that *K* points in the training data that are closest to *x* belong to the set *A*, then the classifier estimates the conditional probability for each class, the fraction of points in A with that given class label:

$$P(y=j \mid X=x)=\frac{1}{K}\sum_{i=A} I\left(y^{(i)}=j\right) \quad (4),$$

where $I(x)$ is the indicator function which evaluates to1 when the argument x is true and 0 otherwise.

Naive Bayes (NB) algorithm

A Naive Bayes classifier is a probabilistic machine learning model that is used for a classification task. The crux of the classifier is based on the Bayes theorem[4].

Bayes Theorem:

$$P(y \mid X)=\frac{P(X \mid y)P(y)}{P(X)} \quad (5)$$

Variable X represent the parameters/features. X is given as,

$$X=(x_1,x_2,x_3,...,x_n) \quad (6)$$

Here $x_1,x_2,x_3,...,x_n$ represent the features. By substituting for X and expanding using the chain rule we get,

$$P(y \mid x_1,x_2,x_3,...,x_n)=\frac{P(x_1 \mid y)P(x_2 \mid y)...P(x_n \mid y)P(y)}{P(x_1)P(x_2)...P(x_n)} \quad (7)$$

Gaussian Naive Bayes:

$$(x_i \mid y)=\frac{1}{\sqrt{2\pi\sigma_y^2}}e^{\left(-\frac{(x_i-\mu y)^2}{2\sigma_y^2}\right)} \quad (8)$$

ADASYN

ADASYN is a method of synthetic data creation for the minority class uses the following function as described by He and Garcia:

$$G=\left(|S_{maj}|-|S_{\min}|\right)\times\beta \quad (9),$$

where $\beta\in[0,1]$ is a parameter used to specify the desired balance level after the synthetic data generation process.

Next, for each example $x_i\in S_{\min}$ find the K-nearest neighbors according to the Euclidean distance and calculate the ratio Γ_i defined as:

$$\Gamma_i=\frac{\Delta_i/K}{Z}, i=1,...,|S_{\min}| \quad (10),$$

where $\varnothing_i$ is the number of examples in the K-nearest neighbors of x_i that belong to S_{maj}, and Z is a normalization constant so that Γ_i is a distribution function $\left(\sum\Gamma_i=1\right)$.

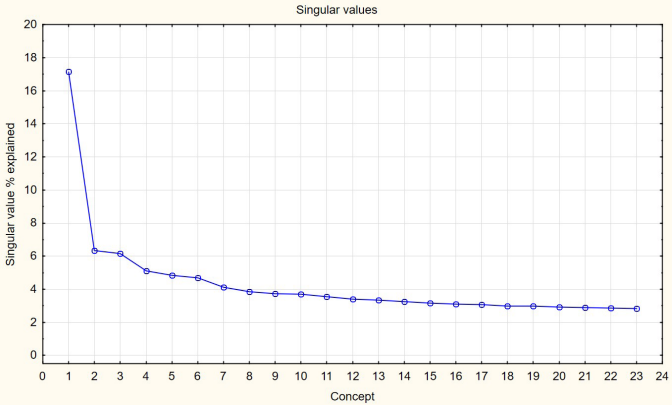
Then, determine the number of synthetic data samples that need to be generated for each $x_i\in S_{\min}: g_i=\Gamma_i\times G$ (11)

Finally, for each $x_i\in S_{\min}$, generate g_i synthetic data samples, according to (7).

The key idea of the ADASYN algorithm is to use a density distribution as a criterion to automatically decide the number of synthetic samples that need to be generated for each minority example by adaptively changing the weights of different minority examples to compensate for the skewed distributions [12].

Analysis and Results

First, we ran a text-mining (TM) procedure to extract the significance of the words from corpus. The concepts visualization of text mining results was as follows (Fig.1):



F i g. 1. TM concepts visualization

Based on the graph above, we identified four significant concepts with greatest variability.

Then we divided the texts into a training (80%) and testing subsets (20%) and trained the machine using the K-NN and NB algorithms. Usually, when forming the original dataset, the data are organized into a file in the sorted or otherwise structured order. To avoid the bias this causes to further ML, we randomized the sample to make cross-validation and divided it into five equal subsets. Then in turns, we used each of these subsets as a testing subset and the remaining documents as the training subset.

The condition of independence between the features implied by the of Bayes' law is partially observed as shown by the low correlations of the four concepts (Fig. 2):

Correlations
Marked correlations are significant at p < ,05000
N=5211 (Casewise deletion of missing data)

Variable	Concept1	Concept2	Concept3	Concept4
Concept1	1,000000	0,217144	-0,371307	0,023517
Concept2	0,217144	1,000000	0,095225	-0,006031
Concept3	-0,371307	0,095225	1,000000	0,010313
Concept4	0,023517	-0,006031	0,010313	1,000000

F i g. 2. TM concepts correlation

ML accuracies and performance

The accuracy percentages obtained for K-NN and NB using Python libraries were as shown in Table 2:

[4] Gandhi R. Naive Bayes Classifier. May 5, 2018. [Electronic recourse]. Available at: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed 10.04.2019). (In Eng.)

Table 2. **Accuracy of K-NN and NB algorithms**

	K-NN	NB
accuracy	0.779482	0.789069
accuracy (%)	78%	79%

To better understand the quality of ML classification, we calculated commonly used performance metrics using Python libraries. Confusion matrices [13] for K-NN and NB algorithms are presented in Tables 3 and four, respectively:

Table 3. **K-NN algorithm confusion matrix**

		negative	neutral	positive
K-NN classification	negative	33	105	1
	neutral	40	777	8
	positive	3	73	3

Table 4. **NB algorithm confusion matrix**

		negative	neutral	positive
NB classification	negative	38	55	6
	neutral	150	654	98
	positive	0	0	0

Other established ML performance metrics - precision-recall, F1 scores [14, 15] for K-NN and NB algorithms are shown in Tables 5 and 6, respectively:

Table 5. **K-NN algorithm performance metrics**

K-NN ML			
precision (neg)=	43%	recall (neg)=	24%
precision (neut)=	81%	recall (neut)=	94%
precision (pos)=	25%	recall (pos)=	4%
precision (macro avg)=	50%	**recall (macro avg)=**	41%
F1 score	42%		

Table 6. **NB algorithm performance metrics**

NB ML			
precision (neg)=	46%	recall (neg)=	33%
precision (neut)=	82%	recall (neut)=	94%
precision (pos)=	0%	recall (pos)=	0%
precision (macro avg)=	43%	**recall (macro avg)=**	42%
F1 score	42%		

The contradiction between accuracies and other performance metrics

From the evidence above we can see a contradiction between acceptable accuracies (~80%) and poor ML performance from the confusion matrices and other measurements calculated for both ML types. In scientific literature, authors commonly report the accuracy scores to prove the quality of their machine learning performance. So here we face the situation where we need some other tool to assess the performance of our ML learning that will allow a better understanding of whether the classification was performed successfully and eliminate the discrepancy between accuracy and other performance metrics.

Acting under the hypothesis that these discrepancies result from data imbalances, we reviewed the literature on dealing with imbalanced data describing established methods like undersampling, simple oversampling, SMOTE and chose ADASYN method as said to be the most appropriate for the datasets similar to the one used in this research [16, 17, 18, 19].

Oversampling with ADASYN

In our case using ADASYN oversampling is beneficial in terms of small sampling, without getting rid of most of the samples (so the model is more prone to error) instead of using synthetic data generation to increase the number of samples in the data set [20]. To address the imbalance, we applied ADASYN oversampling method that generates synthetic data in an adaptive way where the amount of generated data is weighted based on the distributions and the level of difficulty in learning for “positive tone” and “negative tone” minority classes.

According to He et al., the key idea of ADASYN algorithm is to use a density distribution $\hat{r}_i$ as a criterion to automatically decide the number of synthetic samples that need to be generated for each minority data example. Physically, $\hat{r}_i$ is a measurement of the distribution of weights for different minority class examples according to their level of difficulty in learning [12]. The dataset generated by ADASYN will represent a balanced data distribution where balance level is determined adaptively based on the β coefficient, and the ML algorithm applied further will have a particular focus on the minority class examples that are difficult to learn. This principle of differently weighted minority class examples is similar to SMOTEBoost described by Chawla et al. [21] and DataBoost-IM described by Guo and Viktor [22], and the main advantage of ADASYN over the SMOTE [23] algorithm. However, the approach used in ADASYN is more efficient since both SMOTEBoost and DataBoost-IM rely on the evaluation of hypothesis performance to update the distribution function, whereas our algorithm adaptively updates the distribution based on the data distribution characteristics. Hence, there is no hypothesis evaluation required for generating synthetic data samples in our algorithm.

The sets we obtained after ADASYN in Python had the following characteristics (Tab. 7):

Table 7. **Characteristics of the corpus synthetically generated by ADASYN method**

	Count	%
negative	3333	33,24
neutral	3295	32,86
positive	3400	33,91
N total=	10028	100

The confusion matrices for the oversampled with ADASYN datasets and processed with K-NN and NB algorithms are presented in Tables 8 and 9, respectively.

Table 8. **K-NN algorithm confusion matrix post-ADASYN**

		negative	neutral	positive
K-NN classification	negative	67	52	20
	neutral	214	437	174
	positive	12	33	34

Table 9. **NB algorithm confusion matrix post-ADASYN**

		negative	neutral	positive
NB classification	negative	69	31	39
	neutral	100	202	523
	positive	7	3	69

Post ADASYN we observed the following results for K-NN machine learning (Tab.10):

Table 10. **K-NN performance post-ADASYN**

Oversampled by ADASYN - K-NN ML			
accuracy=	52%		
precision (neg)=	23%	recall (neg)=	31%
precision (neut)=	84%	recall (neut)=	65%
precision (pos)=	15%	recall (pos)=	22%
precision (macro avg)=	40%	**recall (macro avg)=**	48%
F1 score	39%		

The results for NB machine learning classification on a corpus generated by ADASYN were as follows (Tab.11).

Table 11. **NB performance post-ADASYN**

Oversampled by ADASYN - NB ML -			
accuracy=	33%		
precision (neg)=	39%	recall (neg)=	50%
precision (neut)=	86%	recall (neut)=	24%
precision (pos)=	11%	recall (pos)=	87%
precision (macro avg)=	45%	**recall (macro avg)=**	34%
F1 score	34%		

After oversampling with ADASYN, we observe the drastic decrease in accuracy and a moderate decrease of F1 scores. Post-synthetic oversampling with ADASYN the discrepancies between accuracy and other performance metrics had leveled down: at that, the difference between accuracy and F1 score comprised 13% and 1% for K-NN and NB algorithms respectively, versus the difference of 36% and 37% for K-NN and NB algorithms before ADASYN.

One important reservation worth mentioning here is that the focus of this paper was not building a model with acceptable performance metrics, but to demonstrate how ADASYN enhances better understanding of data and contributes to the obtainment of the machines learning algorithms performance metrics of better quality more corresponding to the actual state of things.

Discussions

When putting forward the limitations of the presented research work for discussion, it is worth to note that in our case, the machine fails to classify the positive tone texts and exhibit low-quality performance for F1 scores. This misclassification may be explained by the data imbalance which reflects the objective reality of the observed media landscape, where negative toned news prevails over positive toned ones, and the most news produced by news aggregators tend to be neutral and predominate over the other two.

Another aspect that may be considered as a limitation is the size of the dataset as the number of 10 000 texts, and more is commonly considered to be appropriate for text-based machine learning. Nevertheless, most of the papers from scientific journals use much fewer data to conduct research.

When talking about future research work, one area we propose for the future study is comparing the accuracy and precision-recall based performance metrics of the K-NN and NB ML algorithms pre-processed with ADASYN to other methods like SMOTEBoost and DataBoost-IM. Another interesting area for future research may be the study of the effect of ADASYN on other ML performance metrics like ROC curve and AUC, G-mean, optimized precision, Index of Balanced Accuracy, and others.

Conclusion

From the evidence above, we see that if before oversampling, we observed the dramatic difference between the accuracy and precision-recall, F1 metrics, after the oversampling, the difference between the said measurements became evener. This allows concluding that ADASYN may be a useful tool to verify the correspondence between accuracy and precision-recall, F1 metrics and its application to imbalanced corpora may be a way to evaluate the performance of ML on imbalanced corpora deeper.

References

[1] Luque A., Carrasco A., Martin A., Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*. 2019; 91:216-231. (In Eng.) DOI: 10.1016/j.patcog.2019.02.023

[2] Fu G.H., Xu F., Zhang B.Y., Yi L.Zh. Stable variable selection of class-imbalanced data with precision-recall criterion. *Chemometrics and Intelligent Laboratory Systems.* 2017; 171:241-250. (In Eng.) DOI: 10.1016/j.chemolab.2017.10.015

[3] Zhang X., Wang D., Zhou Y., Chen H., Cheng F., Liu M. Kernel modified optimal margin distribution machine for imbalanced data classification. *Pattern Recognition Letters.* 2019; 125:325-332. (In Eng.) DOI: 10.1016/j.patrec.2019.05.005

[4] Haixiang G., Yijing L., Shang J., Mingyun G., Yuanyue H., Bing G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications.* 2017; 73:220-239. (In Eng.) DOI: 10.1016/j.eswa.2016.12.035

[5] Zhang S. Cost-Sensitive KNN Classification. *Neurocomputing*. 2019. (In Eng.) DOI: 10.1016/j.neucom.2018.11.101

[6] Zhang Y.C., Sakhanenko L. The naive Bayes classifier for functional data. *Statistics & Probability Letters.* 2019; 152:137-146. (In Eng.) DOI: 10.1016/j.spl.2019.04.017

[7] Abraham B., Nair M.S. Computer-aided diagnosis of clinically significant prostate cancer from MRI images using sparse autoencoder and random forest classifier. *Biocybernetics and Biomedical Engineering.* 2018; 38(3):733-744. (In Eng.) DOI: 10.1016/j.bbe.2018.06.009

[8] Feeney M. What Can Text Mining Reveal about the Use of Newspapers in Research? *IFLA International News Media Conference "Collecting, Preserving, and Transforming the News – for Research and the Public" 27-28 April 2017.* Landsbókasafn Íslands-Háskólabókasafn (The National and University Library of Iceland), Reykjavik, Iceland, 2017. Available at: https://www.ifla.org/files/assets/newspapers/2017_Iceland/2017-feeney-en.pdf (accessed 10.04.2019). (In Eng.)

[9] Kim D., Seo D., Cho S., Kang P. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Information Sciences.* 2019; 477:15-29. (In Eng.) DOI: 10.1016/j.ins.2018.10.006

[10] Manning C.D., Schutze H. Foundations of Statistical Natural Language Processing. MIT Press, 2000. 620 p. (In Eng.)

[11] Gao J., Zhang J. Clustered SVD strategies in latent semantic indexing. *Information Processing & Management.* 2005; 41(5):1051-1063. (In Eng.) DOI: 10.1016/j.ipm.2004.10.005

[12] He H., & Garcia E.A. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering.* 2009; 21(9):1263-1284. (In Eng.) DOI: 10.1109/tkde.2008.239

[13] Visa S. & Ramsay B., Ralescu A., & Knaap E. Confusion Matrix-based Feature Selection. *CEUR Workshop Proceedings.* 2011; 710:120-127. Available at: http://ceur-ws.org/Vol-710/paper37.pdf (accessed 10.04.2019). (In Eng.)

[14] Prasath S., Arafat A.A.H., Lasassmeh O., & Hassanat A. Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier - A Review. *CoRR.* 2017; abs/1708.04321. Available at: http://arxiv.org/abs/1708.04321 (accessed 10.04.2019). (In Eng.)

[15] Chomboon K., Chujai P., Teerarassammee P., Kerdprasop K., & Kerdprasop N. An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm. *Proceedings of the 3rd International Conference on Industrial Application Engineering 2015.* The Institute of Industrial Applications Engineers, Japan, 2015; p. 280-285. (In Eng.) DOI: 10.12792/iciae2015.051

[16] Picek S., Heuser A., Jovic A., Bhasin S., Regazzoni F. The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-channel Evaluations. *IACR Transactions on Cryptographic Hardware and Embedded Systems.* 2018; 2019(1):209-237. (In Eng.) DOI: 10.13154/tches.v2019.i1.209-237

[17] Ah-Pine J., Soriano-Morales E.P. A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis. *Workshop on Interactions between Data Mining and Natural Language Processing (DMNLP 2016)*, Sep 2016, Riva del Garda, Italy, 2016. ffhal-01504684f. Available at: https://hal.archives-ouvertes.fr/hal-01504684 (accessed 10.04.2019). (In Eng.)

[18] Boughorbel S., Jarray F., El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE*. 2017; 12(6):e0177678. (In Eng.) DOI: 10.1371/journal.pone.0177678.

[19] Maldonado S., Lopez J., Vairetti C. An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing*. 2019; 76:380-389. (In Eng.) DOI: 10.1016/j.asoc.2018.12.024

[20] Chawla N., Japkowicz N., & Kotcz A. Editorial: Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets.* 2004; 6(1):1-6. (In Eng.) DOI: 10.1145/1007730.1007733

[21] Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Lavrač N., Gamberger D., Todorovski L., Blockeel H. (eds). *Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2003; 2838:107-119. (In Eng.) DOI: 10.1007/978-3-540-39804-2_12

[22] Guo H., Viktor H.L. Learning from Imbalanced Data Sets with Boosting and Data Generation: the DataBoost-IM Approach. *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets.* 2004; 6(1):30-39. (In Eng.) DOI: 10.1145/1007730.1007736

[23] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*. 2002; 16: 321-357. (In Eng.) DOI: 10.1613/jair.953

Submitted 10.04.2019; revised 14.05.2019;
published online 25.07.2019.
Поступила 10.04.2019; принята к публикации 14.05.2019;
опубликована онлайн 25.07.2019.

About the authors:

Mukhit M. Baimakhanbetov, Chief Analyst, Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., BC "Moskva", Floor 10, Nur-Sultan 010000, Kazakhstan), ORCID: http://orcid.org/0000-0003-4978-2704, m.baimakhnbetov@gmail.com

Kaidar S. Nurumov, Manager, Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., BC "Moskva", Floor 10, Nur-Sultan 010000, Kazakhstan), Master of Political Science, ORCID: http://orcid.org/0000-0002-1514-0095, k.nurumov@gmail.com

Ulzhan A. Ospanova, Project Manager of the Department of Applied Research and Development, , Information-Analytical Center, JSC (18 Dostyk Ave., BC "Moskva", Floor 10, Nur-Sultan 010000, Kazakhstan), Master of Management, ORCID: http://orcid.org/0000-0003-0427-7931, ulzhan.ospanova@iac.kz

Timur K. Buldybayev, Director of the Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., BC "Moskva", Floor 10, Nur-Sultan 010000, Kazakhstan), ORCID: http://orcid.org/0000-0001-8199-653X, timur.buldybayev@iac.kz

Inessa G. Akoyeva, Chief Analyst of the Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., BC "Moskva", Floor 10, Nur-Sultan 010000, Kazakhstan), ORCID: http://orcid.org/0000-0002-1984-150X, inessa.akoyeva@iac.kz

All authors have read and approved the final manuscript.

Об авторах:

Баймаганбетов Мухит Абилкасымович, главный аналитик Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Казахстан, г. Нур-Султан, пр. Достык, д. 18, БЦ Москва, 10 этаж), ORCID: http://orcid.org/0000-0003-4978-2704, m.baimakhnbetov@gmail.com

Нурумов Кайдар Серикович, менеджер Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Казахстан, г. Нур-Султан, пр. Достык, д. 18, БЦ Москва, 10 этаж), ORCID: http://orcid.org/0000-0002-1514-0095, k.nurumov@gmail.com

Оспанова Улжан Абаевна, проектный менеджер Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Казахстан, г. Нур-Султан, пр. Достык, д. 18, БЦ Москва, 10 этаж), магистр менеджмента, ORCID: http://orcid.org/0000-0003-0427-7931, ulzhan.ospanova@iac.kz

Булдыбаев Тимур Керимбекович, директор Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК 010000, Казахстан, г. Нур-Султан, пр. Достык, д. 18, БЦ Москва, 10 этаж), http://orcid.org/0000-0001-8199-653X, timur.buldybayev@iac.kz

Акоева Инесса Георгиевна, главный аналитик Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Казахстан, г. Нур-Султан, пр. Достык, д. 18, БЦ Москва, 10 этаж), магистр менеджмента, ORCID: http://orcid.org/0000-0002-1984-150X, inessa.akoyeva@iac.kz

Все авторы прочитали и одобрили окончательный вариант рукописи.