ISNN:2181-1296



SHAROF RASHIDOV NOMIDAGI SAMARQAND DAVLAT UNIVERSITETI

САМАРКАНДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ ШАРОФА РАШИДОВА

SAMARKAND STATE UNIVERSITY NAMED AFTER SHAROF RASHIDOV

ILMIY AXBOROTNOMA

НАУЧНЫЙ ВЕСТНИК SCIENTIFIC JOURNAL

ANIQ VA TABIIY FANLAR SERIYASI

Matematika, Mexanika, Informatika
Fizika, Kimyo, Biologiya, Geografiya

№ 3/2 (139)
2023

ISSN 2181-1296

ILMIY AXBOROTNOMA

НАУЧНЫЙ ВЕСТНИК SCIENTIFIC JOURNAL

2023-yil, 3-son (2-seriya)(139) ANIQ VA TABIY FANLAR SERIYASI

Matematika, Mexanika, Informatika, Fizika, Kimyo, Biologiya, Geografiya

Samarqand viloyat matbuot boshqarmasida ro‘yxatdan o‘tish tartibi 09-25.
Jurnal 1999-yildan chop qilina boshlagan va OAK ro‘yxatiga kiritilgan.

BOSH MUHARRIR **R. I. XALMURADOV, t.f.d., professor**
BOSH MUHARRIR O‘RINBOSARLARI: **H.A. XUSHVAQTOV, f-m.f.d., dotsent**
A. M. NASIMOV, t.f.d., professor

TAHRIRIYAT KENGASHI:

ANIQ FANLAR

SH.A.ALIMOV - O‘zFA akademigi
S.N.LAKAYEV - O‘zFA akademigi (SamDU)
M.M.MIRSAIDOV - O‘zFA akademigi
A.S.SOLEEV - f.-m.f.d., professor (SamDU)
I.A.IKROMOV - f.-m.f.d., professor (SamDU)
B.X.XO‘JAYAROV - f.-m.f.d., professor (SamDU)
A.G.YAGOLA - f.-m.f.d., professor (Moskva davlat universiteti, Rossiya)
I.I.JUMANOV - f.-m.f.d., professor (SamDU)
X.X.XUDOYNAZAROV - t.f.d., professor (SamDU)
ALBERTO DEL BIMBO - Florensiya universiteti professori, Italiya
L.SABIROV - f.-m.f.d., professor (SamDU)
A.JUMABOYEV - f.-m.f.d., professor (SamDU)
N.N.NIZAMOV - f.-m.f.d., professor (SamDU)
O.Q.QUVONDIQOV - f.-m.f.d., professor (SamDU)
I.A.RAXMATULLAYEV - f.-m.f.d., professor
A.SH.YARMUXAMEDOV - f.-m.f.n., (SamDU)
X.S.XAYDAROV - f.-m.f.n., dotsent (SamDU)

TABIY FANLAR

M.X.ASHUROV - O‘zFA akademigi
N.B. FERAPONTOV - k.f .d., professor (Moskva davlat universiteti, Rossiya)
SH. M. TUGIZOV - professor, Koliforniya universiteti, AQSh
H. I. AKBAROV - k.f.d., professor (O'zMU)
E. A. ABDURAXMONOV - k.f.d., professor (SamDU)
N. K. MUXAMADIYEV - k.f.d., professor (SamDU)
L. A. BULAVIN - Kiev milliy universiteti professori, Ukraina
X. Q. XAYDAROV - b.f.d., professor (SamDU)
Z. I. IZZATULLAYEV - b.f.d., professor (SamDU)
Sh. T. XOLIQULOV - g.f.d., professor (SamDU)
S. B. ABBASOV - g.f.d., professor (SamDU)
Q. S. YARASHEV - g.f.d., professor (SamDU)
GUN-SIK PARK - Seul univeriteti professori, Koreya
D.B.XURSANOV - g.f.f.d., dotsent (SamDU)
M. S. QUZIYEV - b.f.f.d., dotsent (SamDU)

Obuna indeksi – yakka tartibdagi obunachilar uchun **- 5583,**
tashkilot, korxonalar uchun **- 5584**

MUNDARIJA	
Кулжанов У.Н., Адилов А.А. *Спектралные свойства одночастичного оператора шредингера с контактным потенциалом*……..	3-7
Eshbekov R. *The cauchy problem for the modified korteweg-de vries equation with finite density in the class of periodic functions*……..	8-16
Abdurakhmonov B., Zaripov O., Teshayeva M. *Lagranj ko`paytuvchilari yordamida tengsizliklarni isbotlash*……	17-21
Umirzoqova I., Boboxonova G. *Oshkormas funksiya egriligi va kvadratik formalari*……..	22-26
Qurbonov H.G‘., Rustamov M.K., Gafurova D.A., Temiron N.N., Toshpulatova M.Sh. *Synthesis and properties of difficult combustion material based on modified polyacrylonitrile*……	27-32
Холиков Д.Ш., Худайбердиев З.Б., Исмоилов Х.Ф. *Упрощение нелинейного уравнения крутильных колебаний круглого конического слоя*……	33-38
Kamoliddinov D.U. *Density and regular distribution*……	39-43
Kuliev K., Eshimova M., Turaqulov T. *A new method for norm of a hardy-volterra operator*……	44-50
Kuziev B.N. *A comparative study on feature selection methods in QSAR*……..	51-55
Мардонов Ж.А. *Задача коши для неоднородного лапласова поля*……	56-61
Нуритдинов И., У.Эсанов З., Саидахмедов К.Х., Мажидов Х.М. *Преобразовании Dy1 центров в Dy2 центры в кристаллах* Dy_2SiO_5……	62-66
Akhmadova M.O. *The number of eigenvalues of the discrete schrödinger operator*……	67-71
Курбаниязов А.А., Сырлыбекқызы С., Раманкулов Ш. Ж., Аймбетова И.О., Ещанова Т.Д. *Новые данные высокого временного разрешения о морских и их изменчивости в зависимости от климатических условий*……	72-77
Ikromov I.A., Safarov A.R., Khudoyberdiev D. *Invariant estimates for generalized oscillatory integrals with polynomial phase*……	78-86
Urazbayev A.K. *Amudaryo hozirgi deltasidagi yer usti suv oqimlarining geografik strukturalari*……	87-91
Komilova N.K., Karshibayeva L.K. *Iqlim o‘zgarishlari va aholi kasalliklarining hududiy tarkibi (Jizzax va Sirdaryo viloyatlari misolida)*……	92-99
Нуритдинов И., Эсанов З.У., Саидахмедов К.Х., Ирисов Ш. *Влияние ионизирующего излучения на спектрально-люминесцентные свойства органических сцинтилляторов ups-923a*……	100-104
Alimkulov N.R. *Antropogen landshaftlarni tadqiq etish asosida landshaft-ekologik sharoit tushunchasining shakllanishi*……	105-113
Кувандиков О.К., Усаров У.Т. *Электрические, гальваномагнитные свойства редкого металла при различных температурах*	114-125
Komilova N.K., Ravshanov A.X. *Global iqlim o‘zgarishlari sharoitida O‘zbekistonda yuqumli-parazitar kasalliklar tarqalishining hududiy jihatlari*……	126-132

Рахматуллаев И.А., Ботиров Х.З., Чернега Н.В., Хайдаров Х.С. *Двухфотонно-возбуждаемая люминесценция в полициклических ароматических соединениях.....*	133-138
Anarova Sh.A., Jabbarov J.S., Rustamov A.A. Kattalashtirish masshtabiga bog'liq bo'lmagan fraktal o'lchovlarni geometrik modellashtirish…….	138-142
Nurmuhamedova D.K., Kutlieva G.J., Sohibnazarova X.A., Zairova X.T. *Mahalliy sut achituvchi bakteriyalarning antimikrob va bakteriosinogen faolligini tadqiq etish …….*	143-148
Тайланов Н.А., Уринов С.Х. *Численное моделирование уравнения гросса-питаевского.....*	149-159

UDK 004.89 **DOI:** https://doi.org/10.59251/2181-1296.2023.v3.139.2.2184

A COMPARATIVE STUDY ON FEATURE SELECTION METHODS IN QSAR

B.N.Kuziev
Jizzakh Polytechnic Institute, Associate Professor

Abstract. A quantitative structure-activity relationship (QSAR) relates quantitative chemical structure attributes (molecular descriptors) to a biological activity. QSAR studies have now become attractive in drug discovery and development because their application can save substantial time and human resources. Several parameters are important in the prediction ability of a QSAR model. On the one hand, different statistical methods may be applied to check the linear or nonlinear behavior of a data set. On the other hand, feature selection techniques are applied to decrease the model complexity, to decrease the overfitting/overtraining risk, and to select the most important descriptors from the often more than calculated. The selected descriptors are then linked to a biological activity of the corresponding compound by means of a mathematical model. Different modeling techniques can be applied, some of which explicitly require a feature selection. A QSAR model can be useful in the design of new compounds with improved potency in the class under study. Only molecules with a predicted interesting activity will be synthesized. In the feature selection problem, a learning algorithm is faced with the problem of selecting a relevant subset of features upon which to focus attention, while ignoring the rest. This paper studied the comparative analysis of the Chi-square, Mutual Information, Anova F-value, Fisher Score and SHAP feature selection methods used in QSAR modeling. The Python code written to get experimental results in this article has been uploaded to Github (https://github.com/rifqat/Feature-Selection).

Keywords: approbation, biological activity, biological properties, descriptor, interpretation, criteria, method, model, prediction, regression model, structure, compounds.

QSAR modelida belgilarni tanlash usullarini qiyosiy tadqiq qilish

Annotatsiya. QSAR modelining tarkibiy faolligi molekulyar deskriptorlarning ximiyaviy strukturasini miqdoriy atributlarini ularni biologik faolligi bilan o’zaro bog’laydi. QSAR tadqiqotlari dori vositalarini tadqiq va ixtiro qilishda juda qo’l keladi. QSAR modelining bashorat qilish faoliyatida bir nechta parametrlar muhim hisoblanadi. Bir tomondan belgilardan iborat ma’lumotlar to’plamini chiziqli yoki chiziqli emasligini tekshirish uchun har xil statistik usullar qo’llanilsa, ikkinchi tomondan belgilarni tanlash usullari modelni murakkablik darajasini kamaytirish uchun qo’llanadi. So’ngra tanlab olingan deskriptorlar matematik model yordamida mos birikmani biologik faolligi bilan bog’lanadi. Bunda har xil modellashtirish usullarini qo’llash mumkin bo’lib, ulardan ayrimlari belgilarni tanlashni talab qiladi. QSAR modeli o’rganilayotgan sinfni biologik faolligi yuqori bo’lgan yangi birikmalarni ishlab chiqishda qulay bo’ladi. Bunda faqat avvaldan, biologik faolligi qiziqish uyg’otadigan molekulalar sintez qilinadi. Belgilarni tanlash masalasida algoritm mos belgilar qism to’plamini aniqlash muammosiga duch keladi. Ushbu maqolada QSAR modelida foydalaniladigan Xi-kvadrat, F –qiymat, Fisher ko’rsatkichi va belgilarni tanlashning SHAP usullari qiyosiy tahlil qilingan. Ushbu maqolani eksperiment natijalarining Python kodi Github (https://github.com/rifqat/Feature-Selection) ga yuklab qo’yilgan.

Kalit so’zlar: aprobatsiya, biologik aktivlik, biologik xususiyat, deskriptor, interpretatsiya, kriteriya, metod, model, bashorat, regression model, struktura, aralashma.

Сравнительное исследование методов выбора признаков в QSAR

Аннотация. Количественная взаимосвязь структура-активность (QSAR) связывает количественные атрибуты химической структуры (молекулярные дескрипторы) с биологической активностью. Исследования QSAR теперь стали привлекательными для открытия и разработки лекарств, потому что их применение может сэкономить значительное время и человеческие ресурсы. Несколько параметров важны для прогнозирующей способности модели QSAR. С одной стороны, для проверки линейного или нелинейного поведения набора данных могут применяться различные статистические методы. С другой стороны, методы выбора признаков применяются для снижения сложности модели, снижения риска обучения/переобучения и выбора наиболее важных дескрипторов из более чем

рассчитанных. Затем выбранные дескрипторы связывают с биологической активностью соответствующего соединения с помощью математической модели. Могут применяться различные методы моделирования, некоторые из которых явно требуют выбора признаков. Модель QSAR может быть полезна при разработке новых соединений с улучшенной активностью в изучаемом классе. Будут синтезированы только молекулы с предсказанной интересной активностью. В проблеме выбора признаков алгоритм обучения сталкивается с проблемой выбора соответствующего подмножества признаков, на котором следует сосредоточить внимание, игнорируя при этом остальные. В этой статье изучался сравнительный анализ методов хи-квадрат, взаимной информации, Anova F-значения, показателя Фишера и методов выбора признаков SHAP, используемых в моделировании QSAR. Код Python, написанный для получения экспериментальных результатов в этой статье, был загружен на Github (https://github.com/rifqat/Feature-Selection).

Ключевые слова: апробация, биологическая активность, биологическая свойства, дескриптор, интерпретация, критерия, метод, модель, прогноз, регрессионный модель, структура, соединения.

Introduction. Drug discovery encompasses understanding cellular processes, predicting protein structures, and estimating interactions between a molecule and the normal biological molecular targets [1]. Chemists and biologists would ideally like to fully understand the pathways involved in a disease, and from this knowledge develop a molecule (or several molecules) that can interact with the disease agents to neutralize them. However, many complex interactions are occurring at the cellular level that makes the full rational drug design process extremely difficult. Methods have been developed to circumvent some of these problems through the use of high-throughput screening [1].

The advent of combinational chemistry in the mid-1980s has allowed the synthesis of hundreds, thousands, and even millions of new molecular compounds. The need for a more refined search than simply producing and testing every single molecular combination possible has meant that statistical approaches and, more recently, intelligent computation have become an integral part of the drug production process [2]. Structure—activity relationship (SAR) analysis is one technique used to reduce the search for new drugs. SAR bases its prediction on the assumption that there exists a relationship between the structural or molecular features of a compound and its biological activity (such as chemical activity, aqueous solubility, blood-brain barrier penetration, oral absorption, or toxicity). SAR analysis aims at discovery these rules in order to predict the activity of new molecules based on their physiochemical descriptors [3-4].

Literature review. Machine learning techniques have been successfully applied to SAR analysis to predict if a compound is likely to demonstrate drug-like activity in the presence of a given disease (or simply a given chemical target) [5]. Artificial neural networks have been used to discriminate potential drug-like molecules from large compound databases [3]. Wagener et al.[4] reached 70—80% accuracy in a similar type problem using decision tree algorithms. Burbidge et al. [2] showed that support vector machine outperformed other machine learning approaches for the prediction of inhibition of dihydrofolate reductase by pyrimidines.

Complex molecular compounds can be described by a large number of attributes or features, such as topological indices, characterizing the three-dimensional molecular structures, quantum mechanical descriptors, and molecular field parameters, which could be tens or hundreds of thousands of features. This is prohibitively high for many learning algorithms [6]. Therefore, a first step in the machine learning process consists of identifying the most relevant features for the problem at hand. The larger the number of irrelevant features in the input space is, the more difficult for the algorithms to identify a correct decision function: the system may not converge to an optimal solution in an acceptable amount of time, or much more training data may be needed to| reach a correct solution. To reduce the dimensionality of the input space one will need to (1) identify all the variables relevant to the concept and determine how relevant they are and how related to one another and (2) choose a minimum subset of variables or alternative subsets that maximize the inducer’s efficiency (i.e. provide good generalization). Aggressive reduction of the feature space has been repeatedly shown to lead to little accuracy loss and to a performance gain in text categorization [7].

This paper is to evaluate and compare different feature selection methods in the reduction of a high dimensional feature space in drug discovery. The Logistic Regression method was used to classify the chemical compounds.

Research Methodology. 1. The Data Set. The data used in this paper was from the data set in [8]. Each compound was described by a single feature vector comprised of a class value ("A" for active, "I" for inactive) and 105 features.

2. Feature Selection Methods. In this paper, five methods were evaluated. The methods are all based on assigning a score to each feature that suggests how important or valuable the feature is likely to be for the training and categorization.

Chi-square. This is a univariate filter based on the common 2 statistical test that measures divergence from the expected distribution if one assumes the feature occurrence is actually independent of the class value (Jin et al., 2006). Like any univariate method, we calculate the chi-square between every feature and the target variable, and observe the existence of a relationship between them. If the target variable is independent of the feature, then it gets a low score, or if they are dependent, the feature is important. A higher value of chi-square means that the feature is more relevant concerning the class.

Mutual Information. This method (Quinlan, 1986) is a filter, also called Information Gain (Hoqne et al., 2014). It is one of the most popular and common univariate feature selection methods, because of its computational speed and simple formula. It measures the decrease in entropy considering a single feature at a time. A feature is considered relevant if it has a high information gain. It cannot handle redundant features, because features are selected in a univariate way. So, it belongs to the 'myopic' methods, where features are evaluated independently without considering their context.

Anova F-value. It is a univariate filter method that uses variance to find out the separability of the individual features between classes. It applies to multi-class endpoints (Ding et al., 2014; Jafari and Azuaje, 2006).

Fisher Score. This is a filter method that uses mean and variance to rank the features (Duda et al., 2001). Features with similar values in their instances of the same class and different values to instances from different classes are considered best. Like the previous univariate methods, it evaluates features individually, and it cannot handle feature redundancy (Duda et al., 2012).

SHAP. SHapley Additive exPianations is a unified approach to explain the output of any machine learning model. It connects game theory with local explanations, uniting several previous methods and representing the only possible consistent and locally accurate additive feature attribution method based on expectations (Lundberg and Lee, 2017). It has| become a de-facto feature selection method in the recent years.

Classifiers. After selecting the most discriminatory features, two classifiers were applied to assess the effectiveness of feature selection methods.

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function [https://en.wikipedia.org/wiki/Logistic_function].

As an optimization problem, binary class l_2 penalized logistic regression minimizes the following cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right)$$

Similarly, l_1 regularized logistic regression solves the following optimization problem:

$$\min_{w,c} |w|_1 + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i\left(X_i^T w + c\right)\right) + 1\right).$$

Logistic regression will take in its fit method arrays *X*, *y* and will store the coefficients *w* of the linear model in its coef_ member.

Analysis results. We tested each individual feature selection method and one machine learning algorithms on the one dataset. The experimental procedure is as follows.

1. We performed each individual feature selection method and ranked the features according to their scores.
2. We obtained subsets of the top features (6% for the final run) from all the rankings above.
3. Using the one algorithm described in, we performed classification with a training (80%) and test (20%) fashion. No tuning is performed.
4. We evaluated the results with the Accuracy measure, it can be seen in Table 1.

Table 1. Accuracy Scores for Each Feature Selection Method

Method	Accu-racy	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
Chi-square	0,88	MAC3(Ap)	MAC4(ECI)	MAC4(Ap)	MAC5(Ap)	MAC6(Ap)	MAC7(Ap)
Mutual Information	0,76	MAC4(Ap)	MAC5(Mw)	MAC5(At)	MAC5(Ap)	MAC6(Mw)	MAC6(Ap)
Anova F-value	0,8	MAC4(IP)	MAC4(Ap)	MAC5(IP)	MAC5(At)	MAC5(Ap)	MAC7(Ap)
Fisher Score	0,64	MAC4(ECI)	MAC3(Pa)	MAC3(ECI)	MAC3(Vm)	MAC1(ECI)	MAC4(Pa)
SHAP	0,92	MAC5(Mw)	MAC4(Ap)	MAC4(Z1)	MAC3(Z2)	MAC5(IP)	MAC1(IP)

We used the Python language with the libraries pandas, numpy, and scikit-learn[9] for the main core of our experiments. First, we loaded the datasets and checked for missing values but fomid none. For Chi-square, Mutual Information and Anova F-value, we used the default selector from scikit-learn. For Variance Threshold, we used scikit-learn with the default parameters. For Recursive Feature Elimination, we used scikit-learn and the Logistic Regression algorithm as a black box for the normalized features. This choice was made because it is a popular and relatively fast algorithm with decent results. For SHAP, the shapely package[10] was run with the XGBoost algorithm.

Conclusion. We have experimented with 5 individual feature selection methods. We have tested them for stability and performance on one environmental dataset and one of the strongest classification algorithms in this domain.

Our results indicate that the Reciprocal Ranking outperforms the other methods on performance and also has good stability. The best individual method is SHAP, and another noteworthy is Permutation Importance. Three methods were used for QSAR modeling: multiple linear regression (MLR), random forest (RF), and support vector machine (SVM). The SVD computer program [8] was used to calculate the coefficients of linear equations. A feature of this program is the use of a singular value decomposition of the data matrix, which allows you to effectively identify linear relationships between variables. RF-based regression models were constructed using the original program of the author of the method (Leo Breiman). As controlled parameters, we used such parameters as the number of trees (jbt = 500), the number of randomly selected descriptors (mtry = mdim/3, where mdim is the number of descriptors), the number of objects in the node, below which no tree splitting occurs (nthsize = 5) . The SVM models were calculated using the flssvm program [7]. The compound descriptors were preliminarily autoscaled. The radial basis function was used as the kernel function. The following parameters were used to evaluate the statistical characteristics of QSAR models: n is the number of molecules; m is the number of descriptors; r^2 - is the square of the linear correlation coefficient; rmse is the root mean square error; r_{cv}^2 – is the square of the linear correlation coefficient under cross-control conditions; $rmse_{cv}$ – is the root mean square error under cross-checking conditions; r_{rand}^2 –is the squared randomized correlation coefficient, which was calculated in accordance with the publication [7], using 10 iterations:

$$r_{rand}^2 = (r^2)^{0,5}(r^2 - r_r^2)^{0,5},$$

where r^2 - represents the average value of the square of the linear correlation coefficient for randomized models (those models where the matrix of descriptors remains unchanged, and the dependent variable is used randomly); FIT is a fitness function, which is a modification of the Fisher criterion [8]. When constructing RF models for their testing, we used the out-of-bag procedure, which is an integral part of the algorithm. Given the small size of the initial sample, the test sample was not

isolated from it. Model testing was performed by cross-checking with a choice of 5 and using 10 iterations. As the main strategy for choosing descriptors for constructing models, the method of exhaustive enumeration of combinations of 1, 2, ... 5 descriptors was used. The models were compared by the FIT value of the training set and the model with its maximum value was selected.

Simple SAS cards were used in the work (Fig - 1).

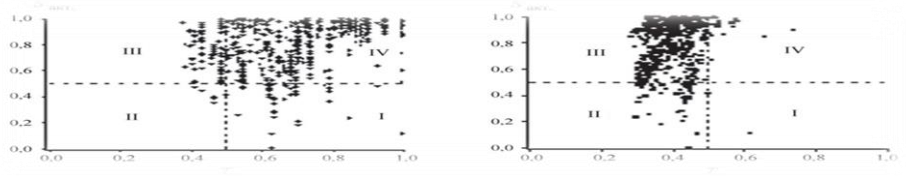

Fig -1.

Each point corresponds to a pair of molecules that have certain values of similarity in structure (T_c) and activity (S_{akt}). Based on the similarity boundaries (dashed lines in Fig. 1), the entire map space can be divided into four areas. In region I, there are molecular pairs that have a high degree of molecular similarity, but differ greatly in activity, activity cliffs (AC). Molecules with low similarity values both in structure and in activity fall into region II, the nondescript region (NR). Region III contains molecular pairs that have a low degree of similarity in structure, but structure cliffs (SC) that are close in activity. Region IV contains molecules that are similar both in structure and activity—smooth region (SR). The main factors influencing the distribution of molecules are:

1) quantitative measures of similarity,
2) method of describing the structure,
3) boundaries of similarity.

In this work, the influence of the first factor was not studied. Two popular types of molecular fingerprints were used to describe the structure of compounds: linear MACCS and spherical ECFP4. As for the boundaries of similarity, different values are given in the literature. Thus, in the cases of MACCS and ECFP4, it is recommended to use T_c values equal to 0.85 and 0.55, respectively [8]. Obviously, these recommendations are empirical in nature. To make the picture more objective, we used a combinatorial approach, which consists in applying nine similarity boundaries in structure and nine similarity boundaries in activity in the range from 0.1 to 0.9 with a step of 0.1, i.e. the total number of combinations of similarity boundaries in structure and activity should be equal to 81. It should be noted that SAS maps contain data on the pairwise distribution of molecules.

References

1. Corwin C., Kuntz I. D. Database searching: Past, present and future. In Designing bioactive molecules', Martin Y. C., Willett. P.. Eds.; American Chemical Society: Washington, DC. 1998: pp 1 — 16.
2. Burbridge R., Trotter M., Buxton B. Holden S. Ding design by machine learning: support vector machines for pharmaceutical data analysis. Proceedings of the AISB'OO Symposium on Artificial Intelligence m Bioinfomiatics, 2000. pp 1 —4.
3. Frimurer T. M.; Bywater R., Namm L., Lauritsen L. N., Bnuiak S. Improving the odds in discriminating drug-like from non drug-like compounds. J. Chem. Inf. Comput. Sci. 2000, 40. 1315 - 1324.
4. Wagener M., van Geerestein V. Potential drugs and nondings: prediction and identification of important structural features. J. Chem. Inf. Comput. Sci. 2000. 40, 280—292.
5. Eschrich S., Chawla N.V., Hall L.O. Generalization methods in Bioinformatics. BIOKDD02: Workshop on Data Mining in Bioin-fomiatics, 2002. pp 25—32.
6. Yang Y., Pederson J. O. A comparative study on feature selection in text categorization. International Conference on machine Leaning (ICML'97), 1997; pp 412-420.
7. Dabrowski Y., Deltorn J.-M. Machine learning application to drug design, http://www.inimngmachines.com.
8. Amit Kumar Halder M. Natália Dias Soeiro Cordeiro QSAR-Co-X: an open source toolkit for multitarget QSAR modelling // Journal of Cheminformatics volume 13, Article number: 29 (2021)
9. https://scikit-learn.org/stable/index.html.
10. https://github.com/Toblerity/Shapely.

SAMARQAND UNIVERSITETI ILMIY AXBOROTNOMASI

НАУЧНЫЙ ВЕСТНИК **SCIENTIFIC REPORTS**

Texnik muharrir **Sh.Xoliqulov**

Muharrirlar:

I.Sulaymonov - fil.f.d., dotsent
E.Arziqulov - f.-m.f.d., dotsent
O.Yusupova - fil.f.n., dotsent

Mas'ul muharrirlar:

A. Sh. Yarmuxamedov - f.-m.f.n.
X. S. Haydarov - f.-m.f.n., dotsent

Muassis: Samarqand davlat universiteti Manzil: 140104,
Samarqand shahri, Universitet xiyoboni, 15.
Telefon: (0 366) 240-38-47, Faks: (0 366) 239-13-87 e-mail:
axborotnoma.uz
SamDU «Ilmiy axborotnoma» jurnali tahririyati kompyuterida terildi. Bosishga 30.06.2023-yilda ruxsat etildi. Qog'oz o'lchami A-4. Nashriyot hisob tabog'i 10,00.
Buyurtma raqami 570. Adadi 30 nusxa.

ISSN 2181-1296

