ENHANCING DRUG DISCOVERY THROUGH A META-PATH BASED OVERSAMPLING APPROACH FOR IMBALANCED DATA

Manh Hung Le[1], *Nam Anh Dao*[1], *Xuan Tho Dang*[2,*]

Abstract

This study proposes a new method to improve the efficiency of drug discovery by repurposing existing drugs, aiming to reduce the time and costs associated with traditional drug development processes, which can span 10 to 15 years and cost billions of dollars. Current approaches focus on leveraging heterogeneous data, such as drug-protein and disease-protein interactions, to construct complex networks that link drugs, proteins, and diseases. However, a significant challenge is the imbalance in data, where numerous unconfirmed potential drug-disease interactions (the majority class) outnumber approved drugs (the minority class), severely impacting the predictive performance of machine learning models. Previous attempts to address this issue have shown limited success. This study introduces a novel approach that integrates meta-paths in heterogeneous information networks with data balancing techniques to tackle this imbalance. Experimental results demonstrate that the proposed method enhances model performance and reliability in identifying new relationships between drugs and diseases. This research represents a promising advancement by leveraging network-based strategies and data balancing techniques to facilitate the rediscovery of drug applications, thereby potentially revolutionizing the pharmaceutical industry's approach to drug development.

Index terms

Drug repositioning, oversampling, undersampling, meta-path, imbalanced data, Gaussian_SMOTE.

1. Introduction

In today's context, the demand for effective drugs to address various diseases, such as infectious diseases, cancer, and rare conditions, is on the rise [1].However, the traditional process of drug discovery and development is both time-consuming and expensive [2]. According to several studies, introducing a new drug to the market typically takes between 10 to 15 years, with an average cost reaching billions of dollars [3]. To address this issue and enhance the efficiency of the drug development process, an increasingly

[1]Electric Power University, Hanoi, Vietnam, [2]Academy of Policy and Development, Hanoi, Vietnam
[*]Corresponding author, email: thodx@apd.edu.vn
DOI: 10.56651/lqdtu.jst.v13.n01.817.ict

prevalent strategy nowadays is drug repurposing, with the expectation of discovering new indications for existing drugs. Based on available data regarding the safety and efficacy of these drugs, this drug repurposing strategy may require less than half the time and investment compared to developing a new drug [3].

By integrating information from two, three, or multiple layers of different data, this method helps to infer potential new relationships between drugs and diseases. Currently, one of the most popular approaches involves using drug-protein interactions, disease-protein interactions, and drug-disease relationships to build a complex drug-protein-disease network [4]–[7].

For most diseases, only a limited number of drugs have undergone clinical validation and are commercially available, representing the positive samples. Conversely, the vast majority of potential drug-disease interactions are unconfirmed and thus classified as negative samples, leading to a significant imbalance within the dataset. Such disparities often impair the efficacy of machine learning models, as they tend to be biased towards the majority class.

Previous research has typically addressed this issue by either randomly selecting negative samples from these unconfirmed drug-disease pairs [8] or by applying rudimentary selection techniques [6], [9] to equilibrate the dataset with positive samples prior to its division into training and testing sets. It is believed that these approaches may drastically reduce the size of the dataset and potentially result in information loss..

A survey has revealed that balancing techniques in drug discovery are not adequately addressed, with only a few studies proposing data balancing methods to enhance model performance [10]–[13]. These studies will be discussed in detail in the related work section of our paper.

This study will attempt to address the imbalance issue and make specific contributions to this end.

a) The proposal to integrate meta-paths in heterogeneous information networks with data balancing techniques aims to boost model performance. This approach is innovative and has not yet been explored in existing research..

b) Experimental results demonstrate that the proposed method effectively tackles the data imbalance problem, thereby enhancing the model's efficiency and increasing the reliability of identifying new drug-disease relationships.

The subsequent sections of this paper are systematically organized as follows: Section 2 evaluates the existing literature, highlighting the pros and cons of previous approaches. Section 3 details the proposed approach, explaining its components and strategy. Section 4 provides a detailed account of experimental results and analyzes the effectiveness of the methodology. Section 5 summarizes key findings, concludes based on the outcomes, and discusses future research directions and advancements.

2. Related work

In recent years, the issue of imbalanced data has garnered significant attention among researchers. Various approaches have been proposed to address the limitations posed by imbalanced data. In this study, we provide an overview of select prior research directly relevant to addressing data imbalance. Imbalance adversely affects the performance of machine learning models. Therefore, addressing this issue has garnered particular attention in various domains such as medical diagnosis [14], fault detection in machinery [15], fraud detection [16], business credit evaluation [17], image recognition [18], etc.

In the field of drug discovery, data balancing techniques have not yet received extensive attention. [19] has employed sampling methods to extract adverse drug reactions from electronic health records, specifically addressing class imbalance issues. [20] developed a model to predict adverse drug events in hospitalized pregnant women using the Synthetic Minority Over-sampling Technique (SMOTE) to enhance the performance of classification algorithms in sparse datasets. [21] used SMOTE to predict inhibitors of the mammalian target of rapamycin in cancer treatment, including the management of asymmetric data through synthetic sample generation for the minority class. [22] analyzed the performance of various resampling techniques in managing imbalanced medical data, utilizing three classical classifiers.

Lastly, [23] proposed a multi-perspective deep neural network model with a novel loss function, "Penalized LF" to address data class imbalance. [24] examined the performance of deep neural network models in classifying imbalanced data (DNNICID), implementing resampling techniques to balance the datasets. [11] introduced a Balanced Matrix Factorization (BMF) method with embedded behavioral information for computational drug repositioning, utilizing a balanced contrastive loss function. Additionally, [12] introduced the method known as RUSBoosted Tree Drug Repurposing (RUSDR). This is a machine learning approach based on ensemble learning, designed to address the issue of class imbalance in datasets. This helps improve model performance on imbalanced datasets, typical of data in drug repurposing where the number of positive drug-disease interactions (positive class) is significantly fewer than the non-positive (negative class). Meanwhile, [13] improved drug-target interaction predictions by addressing both between-class and within-class imbalance with an ensemble learning method (CIEL) that employs decision trees as base learners

Despite the improvements in efficiency brought about by the aforementioned solutions, which partially compensate for the limitations of the data imbalance issue, there are still many constraints. This study proposes enhancing drug discovery with a meta-path based oversampling approach for imbalanced data.

3. The proposed method

In this section, the proposed method is presented and each processing flow is analyzed in detail. Specifically, the framework of the method is illustrated in figure 1. First, a tripartite network among drugs, proteins, and diseases is constructed. Secondly, from

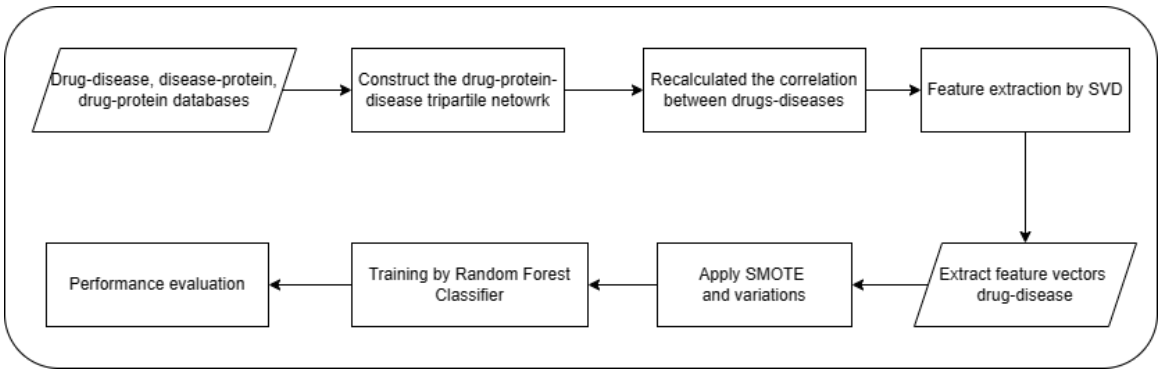


Fig. 1. The workflow consists of five fundamental steps.

the five paths mentioned above, five corresponding drug-disease matrices are built to recompute the correlation between drugs and diseases. Thirdly, due to the large size of the drug-disease vectors, Singular Value Decomposition is employed to extract smaller-sized drug-disease feature vectors. Fourthly, Gaussian_SMOTE techniques are used to balance the data across classes from the aforementioned feature datasets. Finally, the balanced data is fed into a Random Forest (RF) model to construct a classifier for predicting drug-disease relationships.

Construct the drug-protein-disease tripartite network

In this study, drugs are denoted as d, diseases as s, and proteins as p. The corresponding datasets of interactions between drugs-diseases, drugs-proteins, and diseases-proteins are represented by matrices . These datasets are denoted as A^{ds}, A^{dp}, and A^{sp}, respectively, with the number of drugs denoted as n, the number of diseases denoted as m, and the number of proteins denoted as z. A Graph $G = (V, E)$ is defined, where the set of nodes on the network is denoted as V, with each drug (d), disease (s), and protein (p) being respectively represented as a node, denoted as d, s, and p, belonging to V. The edges are denoted as E and are determined from the relationships between the nodes. If a relationship is established between two nodes, an edge is defined between them; otherwise, if there is no known association, there is no edge between the two nodes. To determine the correlation between drugs and diseases, paths from drugs to diseases are constructed. All proposed paths start from any drug and end at a disease, with drug-disease pairs being correlated if there exists a path between them.

To explore the relationship between drugs and diseases, we employed five paths from drugs to diseases as outlined in [6], denoted as M1, M2, M3, M4, and M5, specified as follows:

M1: Represents a direct path from the drug to the disease.

M2: Represents a path from drug-protein-disease. In this case, it reflects the scenario where if a drug and a disease share a protein, there exists a certain degree of similarity between the drug and the disease.

M3: Represents a path from drug-protein-drug-disease. This scenario typically represents the relationship between drugs through proteins, indicating that if two drugs are similar, the diseases they treat may also exhibit similarity.

M4: Represents a path from drug-disease-drug-disease. This case illustrates that if two drugs treat the same disease, they may also have the potential to treat other similar diseases.

M5: Represents a path from drug-disease-protein-disease. In this case, if two diseases share a protein, there is also the possibility that they could be treated by the same type of drug.

Recalculated the correlation between drugs and diseases

In the previous section, five paths representing the routes from drugs to diseases were introduced. Each meta-path provides a flexible and powerful framework for analyzing and modeling complex interactions in heterogeneous networks. It includes both structural characteristics and semantic connections in the network, allowing the extraction of meaningful paths between node pairs. For each path, a relationship matrix between drugs and diseases is constructed, denoted as $A1, A2, A3, A4$, and $A5$, respectively. The correlation between the drug (d_i) and the disease (s_j) in the matrix is computed as follows:

$$A1_{ij} = A_{ij}^{ds} \tag{1}$$

$$A2_{ij} = \sum_{t=1}^{k} A_{it}^{dp} \times A_{tj}^{sp.T} \tag{2}$$

$$A3_{ij} = \sum_{t=1}^{k} \sum_{z=1}^{n} A_{it}^{dp} \times A_{tz}^{dp.T} \times A_{zj}^{ds} \tag{3}$$

$$A4_{ij} = \sum_{t=1}^{m} \sum_{z=1}^{n} A_{it}^{ds} \times A_{tz}^{ds.T} \times A_{zj}^{ds} \tag{4}$$

$$A5_{ij} = \sum_{t=1}^{m} \sum_{z=1}^{k} A_{it}^{ds} \times A_{tz}^{sp} \times A_{zj}^{sp.T} \tag{5}$$

These meta-paths offer diverse perspectives for capturing different aspects of interactions between drugs, proteins, and diseases.By leveraging the structural and semantic information encoded within these meta-paths, effective models can be developed for analyzing and predicting the relationships between diseases and drugs in heterogeneous networks.

Feature extract by Singular Value Decomposition After calculating the correlations between drugs and diseases, five corresponding matrices were obtained: $A1, A2, A3, A4$, and $A5$. Each matrix has dimensions of $n \times m$, representing the correlations between drug-disease pairs in various ways. Due to the large number of drugs and diseases, the size of the drug-disease interaction matrix becomes substantial, leading to increased model complexity and limited generalizability. To address this issue, Singular Value Decomposition was utilized to reduce the dimensions of these matrices. During the dimensionality reduction process, redundant information was discarded while retaining essential data. For each drug-disease interaction matrix, X was obtained by using the top r leading singular values, and row i in U and row j in V were utilized as latent features for the drug (d_i) and disease (s_j), respectively. Specifically, the formula was calculated as follows:

$$X \approx U_{m \times r} \Sigma_{r \times r} V^T_{r \times n} \tag{6}$$

Imbalance and classifiers

The integration of meta-paths in heterogeneous information networks with data balancing techniques is proposed to enhance model performance. This contribution is particularly innovative, as it introduces an approach not previously explored in existing research. Building on this foundation, the study further evaluates and tests a series of data balancing techniques to identify the most effective method for the complex drug-protein-disease network problem. Among these, Gaussian SMOTE [25], an advanced variant of SMOTE, stands out for its efficacy in enhancing model performance. Gaussian SMOTE operates by first identifying disparities between data points in the minority class and their randomly selected nearest neighbors. The technique then applies a Gaussian distribution to these differences, creating diverse synthesized samples. This expansion of the synthetic sampling area significantly improves the reliability of the algorithm by strategically positioning the synthetic data points near the line connecting minority class data points.

Fundamentally, undersampling and oversampling are primary approaches. The simplest of these techniques, Random Undersampling, randomly removes samples from the majority class to balance with the minority class but risks losing important samples. To address this issue, techniques such as Neighbourhood Cleaning Rule (NCR) [26], One-Sided Selection (OSS) [27], Tomek Link [28] ADASYN [29] and SPY [30] have been proposed to optimize the undersampling process by preserving the representativeness of the minority class and retaining important samples.

Other variants of SMOTE, including Borderline SMOTE [31], CURE SMOTE [32], SMOTETomek [33], AND_SMOTE [34], SMOTE_D [35], Random SMOTE [36], SMOTEWB [37], and Kmeans SMOTE [38], were also examined. Each technique offers unique strategies to enhance the diversity of synthetic samples.

As a result, Gaussian SMOTE has proven to be the most effective technique among

those tested, particularly suited for our complex problem. It provides a promising solution for balancing the dataset and improving the model's performance in identifying new drug-disease relationships.

4. Experiments

Data

The data used in this study was extracted from three sources: DrugBank [39], OMIM [10], and Gottlieb [40], corresponding to three sets of drug-disease, drug-protein, and disease- protein relationships. The Drug-Protein dataset encompasses 1,186 drugs and 449 diseases, with a total of 530,687 drug-disease interactions recorded, resulting in a sparsity ratio of 0.334%. The Drug-Protein dataset includes 1,186 drugs, 1,467 diseases, with 1,735,220 pairs identified, achieving a sparsity ratio of 0.267%. The Disease-Protein dataset includes 449 diseases, 1,467 proteins, with 657,318 pairs identified, resulting in a sparsity ratio of 0.207%. We represent the three datasets, drug-disease, drug-protein, and disease-protein, in the form of three matrices denoted as A_{nm}^{ds}, A_{nz}^{dp}, and A_{mz}^{sp}, respectively as presented in table 1. Where $n = 1,186, m = 449, z = 1,467$ are the sizes of the above matrices.

Table 1. Description of experimental data set

Dataset source	Drug	Disease	Protein	Number of no interactions	Ratio in %
OMIN	✓	449	1,467	657,318	0.207
DrugBank	1,186	449	✓	530,687	0.267
Gottlieb	1,186	✓	1,467	1,735,220	0.344

Metrics

In scientific research and data analysis, evaluating the performance of prediction models is of paramount importance. In the context of binary classification with imbalanced data, the minority class is often the primary focus. However, prediction models frequently exhibit a bias toward the majority class. The selection of appropriate parameters for evaluating model performance is crucial.

The confusion matrix, represented in table 2, is commonly used in binary classification to assess model performance. Among the myriad of metrics available for evaluating classification models, each metric is tailored to specific model types and diverse datasets. In the case of severe data imbalance, where the minority class is significantly less prevalent, this study opts for the Geometric mean (G-mean), F1-score, and Precision-Recall Area Under the Curve (PR AUC) as evaluation metrics.

$$F1 - score = \frac{2 \times PRE \times REC}{PRE + REC} \tag{7}$$

Table 2. A confusion matrix for binary class classification

	Predicted Positive	Predicted Negative
Observed Positive	TP	FN
Observed Negative	FP	TN

Here:

Precision (PRE) is calculated as $PRE = TP/(TP + FP)$

Recall (REC) is calculated as $REC = TP/(TP + FN)$

$$G - mean = \sqrt{\text{SP} \times \text{SE}} \tag{8}$$

Here:

Sensitivity (SP) is calculated as $SP = TP/(TP + FN)$

Specificity (SE) is calculated as $SP = TN/(FP + TN)$

In the context of imbalanced data, PR AUC becomes particularly relevant because it provides a more accurate assessment of a model's performance in detecting positive cases (which are often the crucial instances) compared to the area under the ROC curve. This is due to the fact that in imbalanced data, precision and recall are often more critical than the False Positive Rate (FPR) and True Negative Rate (TNR) measured by the ROC Curve. PR AUC is a valuable metric for evaluating the classification model's performance, especially in the context of imbalanced class distribution.

Discussion

First, experiments were conducted on various variants of the SMOTE, including SPY, OSS, NCR, Borderline_SMOTE, CURE_SMOTE, SMOTE_Tomek, and Gaussian_SMOTE. For each method, 5-fold cross-validation was performed. The performance results of each method are presented in table 3.

Table 2 compares the performance of Gaussian_SMOTE with other data balancing techniques. Kmeans_SMOTE is noteworthy with an F1 score of 79.46% and a PR_AUC of 83.45%. Gaussian_SMOTE exhibits superior performance in terms of G-mean at 87.20%, with F1 and PR_AUC scores of 79.37% and 83.39%, respectively. The performance difference between Gaussian_SMOTE and Kmeans_SMOTE is not significant. Additionally, these findings are illustrated in figure 2.

As analyzed above, for imbalanced data, the evaluation metric of interest is often the G-mean. This metric provides a balanced evaluation by considering both sensitivity and specificity, which is crucial for imbalanced datasets. Therefore, in this study, we chose Gaussian_SMOTE as the primary technique due to its superior G-mean performance.

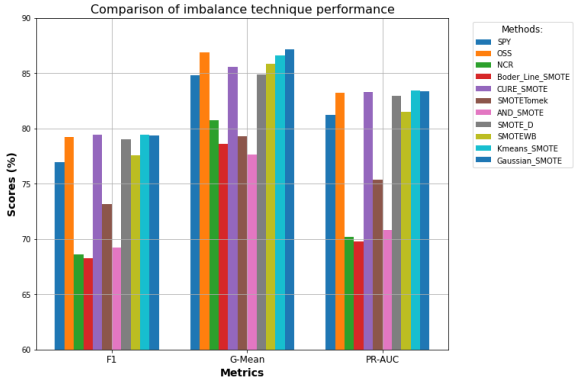


Fig. 2. Comparison of model performance using different data balancing techniques.

This choice is supported by its ability to improve the model's ability to correctly identify both minority and majority class samples, making it a robust method for handling imbalanced data.

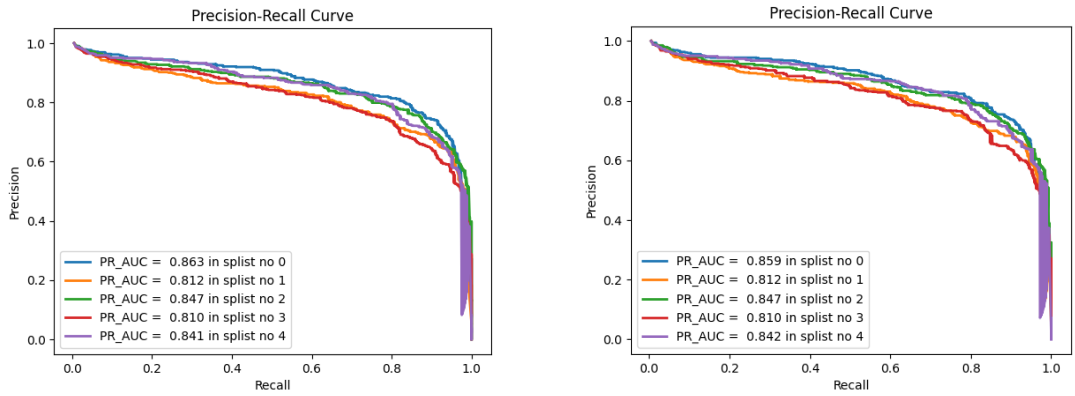
Table 3. Performance of related methods

Method*	F1	G_mean	PR_AUC
SPY	76.98%	84.80%	81.20%
OSS	79.26%	86.86%	83.21%
NCR	68.63%	80.75%	70.17%
Boder_Line_SMOTE	68.26%	78.62%	69.81%
CURE_SMOTE	79.43%	85.58%	83.33%
SMOTETomek	73.16%	79.28%	75.4%
AND_SMOTE	69.21%	77.67%	70.81%
SMOTE_D	79.05%	84.86%	82.93%
SMOTEWB	77.6%	85.85%	81.52%
Kmeans_SMOTE	**79.46%**	86.64%	8**3.45%**
Gaussian_SMOTE	79.37%	**87.20%**	83.39%

∗ the best scores are printed in bold.

The PR_AUC is a performance measurement method for classifying models, especially useful when there is an imbalance between classes. PR_AUC provides an overall view of the model's capability to identify samples belonging to the positive class at different decision thresholds, aiding in adjusting the model's sensitivity and precision. Among the experimented variations of SMOTE models, Kmeans_SMOTE and Gaussian_SMOTE stand out with G_mean scores of 86.64% and 87.29%, respectively. The PR_AUC graphs for these models are depicted in figures 3a and 3b. Notably, both models demonstrate superior G_mean results, with Gaussian_SMOTE exhibiting particularly impressive performance.

As reviewed in section II. Related work, there has been an explosion of research on drug repurposing in recent years. Many of these studies have highlighted the challenges



a. KMeans_SMOTE_PR_AUC *b. Gaussian_SMOTE_PR_AUC*

Fig. 3. ROC Curve Using KMeans_SMOTE and Gaussian_SMOTE Techniques for Imbalanced Data.

of class imbalance in machine learning datasets. For instance, Korkmaz et al. [24] utilized data from the PubChem repository and applied one undersampling technique, Random Undersampling, and three oversampling techniques: SMOTE, ADASYN, and ROS, to enhance model accuracy. Next, Seyede et al. [12] employed the PREDICT Dataset and applied the RUSBoost technique, which combines random undersampling with boosting to handle class imbalance in drug repurposing datasets.

Table 4. Comparison of the proposed method with similar methods in the drug repurposing problem

Techniques*	F1	G_mean	PR_AUC
Seyede et al. [12]	64.12%	76.97%	62.91%
Korkmaz et al. [24]	76.40%	85.68%	80.46%
Our method	**79.37%**	**87.20%**	**83.39%**

∗ the best scores are printed in bold.

In this study, the techniques from [24] and [12] were also applied to the dataset for comparison. The detailed results are presented in table 4. Table 4 demonstrates that the method outperforms the techniques used in [24] and [12] across all three metrics: G-mean, F1, and PR_AUC. Specifically, the method achieves a high G-mean of 87.20%, indicating a better balance between sensitivity and specificity.

Case study

In this section, experiments were conducted on patients diagnosed with Acute Myeloid Leukemia (AML) with the ID 601626. All correlations between AML and various drug categories in the experimental dataset were systematically set to 0. The top 10 predicted outcomes of the model were extracted and presented in the table 5. From table 5, it is evident that 5 drugs have been identified for treating the disease, while 5 newly predicted drugs have emerged.

Through meticulous internet searches for the newly predicted drugs, two of them, Cisplatin and Fludarabine, have shown potential in AML treatment. Mody et al. [41] substantiated the latent therapeutic effects of cisplatin on AML in a patient with a dual diagnosis. Additionally, Carella et al. [42] emphasized the synergistic combination of fludarabine and cytarabine as a preferred approach for AML patients.

Table 5. Top 10 predict drug for disease Acute myeloid leukemia

Drug ID	Drug Name	Known before	Literature validation
DB00541	Vincristine	1	1
DB00997	Doxorubicin	1	1
DB00290	Bleomycin	0	0
DB00694	Daunorubicin	1	1
DB01204	Mitoxantrone	1	1
DB01177	Idarubicin	1	1
DB01033	Mercaptopurine	0	0
DB00515	Cisplatin	0	Mody [41]
DB01073	Fludarabine	0	Carella [42]
DB00444	Teniposide	0	0

5. Conclusions

In this paper, a model based on addressing data imbalance for drug repositioning is proposed. The objective is to enhance predictions of new drugs for disease treatment, thereby reducing time and costs in the pharmaceutical industry. To improve the prediction capabilities of drug-disease associations, five meta-paths are combined with the Gaussian_SMOTE data balancing method. Beyond demonstrating superiority through metrics, the method successfully predicted two drugs for the treatment of AML. This approach presents an appealing option in the pharmaceutical manufacturing field.

References

[1] J. Li, S. Zheng, B. Chen, A. Butte, S. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning. briefings in bioinformatics," *Brief Bioinform*, vol. 17, no. 1, pp. 2–12, 2016. doi: 10.1093/bib/bbv020

[2] G. Law, J. Tisoncik-Go, M. Korth, and M. Katze, "Drug repurposing: a better approach for infectious disease drug discovery?" *Curr Opin Immunol*, vol. 25, no. 5, pp. 588–592, 2013. doi: 10.1016/j.coi.2013.08.004

[3] B. S. Nelson, D. M. Kremer, and C. A. Lyssiotis, "New tricks for an old drug," *Nature Chemical Biology*, vol. 14, no. 11, pp. 990–991, 2018. doi: 10.1038/s41589-018-0137-x

[4] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Information Fusion*, vol. 50, pp. 71–91, 2019. doi: 10.1016/j.inffus.2018.09.012

[5] X. T. Dang, M. H. Le, and N. A. Dao, "Drug repositioning for drug disease association in meta-paths," *Deep Learning and Other Soft Computing Techniques: Biomedical and Related Applications*, pp. 39–51, 2023.

[6] G. Wu, J. Liu, and X. Yue, "Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition," *BMC Bioinformatics*, vol. 20, no. Suppl 3, p. 134, 2019. doi: 10.1186/s12859-019-2644-5

[7] A. Fernández-Torras, M. Duran-Frigola, M. Bertoni, M. Locatelli, and P. Aloy, "Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque," *Nature Communications*, vol. 13, no. 1, p. 5304, 2022. doi: 10.1038/s41467-022-33026-0

[8] L. Wang, Z. H. You, X. Chen, ..., and K. J. Song, "A computational-based method for predicting drug-target interactions by using stacked autoencoder deep neural network," *Journal of Computational Biology*, vol. 25, no. 3, pp. 361–373, Mar 2018. doi: 10.1089/cmb.2017.0135

[9] Y. Huang, Z. You, and X. Chen, "A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences," *Current Protein and Peptide Science*, vol. 19, no. 5, pp. 468–478, 2018. doi: 10.2174/1389203718666161122103057

[10] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, and V. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 33, no. 1, pp. D514–D517, Jan 2005. doi: 10.1093/nar/gki033

[11] X. Yang, G. Yang, and C. J., "The balanced matrix factorization for computational drug repositioning," *arXiv*, 2023. doi: 10.48550/arXiv.2301.06448

[12] S. S. Seyedeh and R. K. Mohammad, "RUSDR: Class imbalance-aware ensemble learning for drug repurposing," *in 10th International Conference on Information and Knowledge Technology (IKT 2019)*, 2019.

[13] A. Ezzat, M. Wu, X. Li, and et al., "Drug-target interaction prediction via class imbalance-aware ensemble learning," *BMC Bioinformatics*, vol. 17, no. Suppl 19, p. 509, 2016. doi: 10.1186/s12859-016-1377-y

[14] R. Rao, S. Krishnan, and R. Niculescu, "Data mining for improved cardiac care," *ACM SIGKDD Explorations Newsletter*, vol. 8, no. 1, pp. 3–10, 2006. doi: 10.1145/1147234.1147236

[15] K. Kerdprasop and N. Kerdprasop, "Data preparation techniques for improving rare class prediction 2 intelligent methods for predicting," *in Proceedings of the 13th WSEAS international conference on mathematical methods*, p. 204–209, 2011.

[16] M. P. Jesús, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, and M. I. José, "Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance," *Lect. Notes Comput. Sci.*, p. 381–389, 2005. doi: 10.1007/11551188_41

[17] L. U. Wang and C. Wu, "Dynamic imbalanced business credit evaluation based on Learn++ with sliding time window and weight sampling and FCM with multiple kernels," *Inf. Sci.*, vol. 520, p. 305–323, 2020. doi: 10.1016/j.ins.2020.02.011

[18] L. Daniel, R. V. Kevelaer, and T. W. Nattkemper, "Strategies for tackling the class imbalance problem in marine image classification," *Int. Conf. on Pattern Recognition*, vol. 2018, p. 26–36, 2018. doi: 10.1007/978-3-030-05792-3_3

[19] S. Santiso, A. Casillas, and A. Pérez, "The class imbalance problem detecting adverse drug reactions in electronic health records," *Health Informatics Journal*, vol. 25, no. 4, pp. 1768–1778, 2019. doi: 10.1177/1460458218799470

[20] L. M. Taft, R. S. Evans, C. R. Shyu, and et al., "Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery," *J Biomed Inform.*, vol. 42, no. 2, pp. 356–364, 2009. doi: 10.1016/j.jbi.2008.09.001

[21] C. Kumari, M. Abulaish, and N. Subbarao, "Using SMOTE to deal with class-imbalance problem in bioactivity data to predict mtor inhibitors," *SN COMPUT. SCI.*, vol. 1, p. 150, 2020. doi: 10.1007/s42979-020-00156-5

[22] M. Khushi and et al., "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021. doi: 10.1109/ACCESS.2021.3102399

[23] S. Mitra, S. Saha, and M. Hasanuzzaman, "A multi-view deep neural network model for chemical-disease relation extraction from imbalanced datasets," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3315–3325, Nov 2020. doi: 10.1109/JBHI.2020.2983365

[24] S. Korkmaz, "Deep learning-based imbalanced data classification for drug discovery," *Journal of Chemical Information and Modeling*, vol. 60, no. 9, pp. 4180–4190, 2020. doi: 10.1021/acs.jcim.9b01162

[25] L. Hansoo, K. Jonggeun, and K. Sungshin, "Gaussian-based SMOTE algorithm for solving skewed class distributions," *Int. J. Fuzzy Logic and Intelligent Systems*, pp. 229–234, 2017. doi: 10.5391/IJFIS.2017.17.4.229

[26] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," *in Proceedings of the 14th International Conference on Machine Learning*, pp. 179–186, 1997.

[27] H. Haibo and M. Yunqian, "Imbalanced learning: Foundations, algorithms, and applications," *Wiley-IEEE Press*, 2013.

[28] I. Tomek, "Two modifications of CNN," *In Systems, Man, and Cybernetics, IEEE Transactions on*, no. 6, pp. 769–772. doi: 10.1109/TSMC.1976.4309452

[29] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," *in Proceedings of IJCNN*, pp. 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969

[30] X. T. Dang, D. H. Tran, O. Hirose, and K. Satou, "SPY: A novel resampling method for improving classification performance in imbalanced data," *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pp. 280–285, 2015. doi: 10.1109/KSE.2015.24

[31] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *in Advances in Intelligent Computing. ICIC. Lecture Notes in Computer Science*, vol. 3644, p. 878–887, 2005.

[32] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, p. 169, 2017. doi: 10.1186/s12859-017-1578-z

[33] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, pp. 20–29, 2004. doi: 10.1145/1007730.1007735

[34] J. Yun, J. Ha, and J. S. Lee, "Automatic determination of neighborhood size in SMOTE," *in Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, vol. 100, pp. 1–8, 2016. doi: 10.1145/2857546.2857648

[35] F. R. Torres, J. A. Carrasco-Ochoa, and J. Martínez-Trinidad, "SMOTE-d a deterministic version of SMOTE," *Pattern Recognition. MCPR 2016. Lecture Notes in Computer Science*, vol. 9703, 2016. doi: 10.1007/978-3-319-39393-3_18

[36] Y. Dong and X. Wang, "A new over-sampling approach: Random-SMOTE for learning from imbalanced data sets," *in Knowledge Science, Engineering and Management. KSEM 2011. Lecture Notes in Computer Science*, vol. 7091, 2011. doi: 10.1007/978-3-642-25975-3_30

[37] F. Sağlam and M. A. Cengiz, "A novel SMOTE-based resampling technique trough noise detection and the boosting procedure," *Expert Systems with Applications*, p. 117023, 2022. doi: 10.1016/j.eswa.2022.117023

[38] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, pp. 1–20, 2018. doi: 10.1016/j.ins.2018.06.056

[39] D. Wishart, Y. Feunang, A. Guo, ..., and M. Wilson, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2018. doi: 10.1093/nar/gkx1037

[40] A. Gottlieb, G. Stein, E. Ruppin, and R. Sharan, "Predict: a method for inferring novel drug indications with application to personalized medicine," *Mol Syst Biol.*, vol. 7, p. 496, 2011. doi: 10.1038/msb.2011.26

[41] M. D. Mody, H. S. Gill, K. A. Higgins, N. F. Saba, and V. K. Kota, "Complete remission of acute myeloid leukemia following cisplatin based concurrent therapy with radiation for squamous cell laryngeal cancer," *Case Reports in Hematology*, vol. 2016, p. 8581421, 2016. doi: 10.1155/2016/8581421

[42] A. M. Carella, N. Cascavilla, M. M. Greco, L. Melillo, and M. Carotenuto, "Treatment of "poor risk" Acute Myeloid Leukemia with Fludarabine, Cytarabine and G-CSF (flag regimen): A single center study," *Leukemia Lymphoma*, vol. 40, no. 3-4, pp. 295–303, 2001. doi: 10.3109/10428190109057928

Manuscript received 02-01-2024; Accepted 25-06-2024.

■

Manh Hung Le obtained a Bachelor's degree from Hong Duc University and a Master's degree in Imformation Technology from the Posts and Telecommunications Institute of Technology in Hanoi, Vietnam. He currently serves as a lecturer in the Faculty of Information Technology at Electric Power University, Hanoi, Vietnam. His research interests encompass various areas, including Bio data processing and analysis, Pattern Recognition, Natural Language Processing, Expert Systems, Machine Learning, Blockchain, and Big Data. Specifically, within the realm of bioinformatics and computational biology, his focus extends to precision medicine, genomics, gait recognition, and public health. E-mail: hunglm@epu.edu.vn

Nam Anh Dao received the B.S degree in applied mathematics and the Ph.D. degree in physics-mathe-matics from the University of Moldova, in 1987 and 1992, respectively. He was involwed in various tesaching at Electric Power University. His reseach interests include intellectual intelligence, image processing and pattern recognition, machine vision, and data science. His main works cover pattern recognition and image analysis, medical imaging, and machine learning with emphasis on computer vision. He also served or is currently serving as a reviewer for many important journals and conferenced in image processing and pattern re. E-mail: anhdn@epu.edu.vn

Xuan Tho Dang has obtained his Bachelor's degree in Information Technology and Master's degree in Computer Science from the Faculty of Information Technology, Hanoi National University of Education, in 2007 and 2009, respectively. He further received his Doctorate in Computer Science from Kanazawa University, Japan, in 2013. Currently, he serves as a lecturer in the Faculty of Digital Economics at the Academy of Policy and Development. His primary research areas focus on machine learning, artificial intelligence, and data mining.
E-mail: thodx@apd.edu.vn

TĂNG CƯỜNG KHÁM PHÁ THUỐC THÔNG QUA PHƯƠNG PHÁP LẤY MẪU QUÁ MỨC DỰA TRÊN SIÊU ĐƯỜNG DẪN CHO DỮ LIỆU KHÔNG CÂN BẰNG

Lê Mạnh Hùng, Đào Nam Anh, Đặng Xuân Thọ

Tóm tắt

Nghiên cứu này đề xuất một phương pháp mới nhằm nâng cao hiệu quả nghiên cứu thuốc bằng cách tái sử dụng các loại thuốc hiện có, nhằm giảm thời gian và chi phí liên quan đến quá trình phát triển thuốc truyền thống, có thể kéo dài từ 10 đến 15 năm và tiêu tốn hàng tỷ đô la. Các phương pháp tiếp cận hiện tại tập trung vào việc tận dụng dữ liệu không đồng nhất, chẳng hạn như tương tác thuốc-protein và bệnh-protein, để xây dựng các mạng lưới phức tạp liên kết thuốc, protein và bệnh tật. Tuy nhiên, một thách thức đáng kể là sự mất cân bằng dữ liệu, trong đó số lượng tương tác giữa thuốc và bệnh tiềm ẩn chưa được xác nhận (nhóm đa số) nhiều hơn số lượng thuốc được phê duyệt (nhóm thiểu số), ảnh hưởng nghiêm trọng đến hiệu suất dự đoán của các mô hình học máy. Những nỗ lực trước đây để giải quyết vấn đề này đã cho thấy thành công hạn chế. Nghiên cứu này giới thiệu một cách tiếp cận mới tích hợp các đường dẫn meta trong các mạng thông tin không đồng nhất với các kỹ thuật cân bằng dữ liệu để giải quyết sự mất cân bằng này. Kết quả thực nghiệm chứng minh rằng phương pháp đề xuất giúp nâng cao hiệu suất và độ tin cậy của mô hình trong việc xác định mối quan hệ mới giữa thuốc và bệnh tật. Nghiên cứu này thể hiện một tiến bộ đầy hứa hẹn bằng cách tận dụng các chiến lược dựa trên mạng và kỹ thuật cân bằng dữ liệu để tạo điều kiện thuận lợi cho việc khám phá lại các ứng dụng thuốc, từ đó có khả năng cách mạng hóa cách tiếp cận phát triển thuốc của ngành dược phẩm.

Từ khóa

Tái định vị thuốc, lấy mẫu quá mức, lấy mẫu dưới mức, siêu đường dẫn, mất cân bằng, Gaussian_SMOTE.