IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTING POST-OPLENGTH OF STAY IN LUNG CANCER WITH SMOTE

THUMMA MANOJNA REDDY

STUDENT

WOXSEN UNIVERSITY

Abstract:

Lung cancer is a frequent and fatal disease that requires state-of-the-art techniques for estimating postoperative survival. This research aims to predict the 3 to 5 years survival of lung cancer patients after surgery by bridging the gap between data-driven methodologies and medical decision support. The research uses enhanced Synthetic Minority Over-sampling Technique (SMOTE) and Fuzzy Particle Swarm Optimization Support Vector Machine (FPSO-SVM) to address class imbalance and SVM parameter optimization difficulties. Using SMOTE for data balance and Cross-Validation Clustering and Filtering (CVCF) to remove noisy data, the dataset is cleaned up in the first stage. The second stage's survival is predicted by the FPSO-SVM model. This innovative method has the potential to enhance medical diagnosis, especially when the datasets are uneven.

Introduction:

The most common cancer that results in mortality is lung cancer, which develops as a malignant tumour in the lungs [1]. Annually, 2.2 million new cases of lung cancer were identified worldwide. The reported incidence rates of lung cancer could be impacted by improvements in early detection techniques, such as low-dose CT scans, which could help diagnose the disease earlier. Selecting candidates for surgery based on the patient's short- and long-term risks and benefits, where survival time is one of the most important indicators, is a significant challenge in the clinical choice on LC procedure. Medical professionals can choose more effective

treatments if they can accurately anticipate a patient's survival following surgery [2].

The postoperative survival of LCPs has recently been predicted using increasingly data-driven techniques. The three most used statistical techniques for predicting survival or complications for LCPs are Kaplan-Meier angles, multivariable logistic retrogression, and Cox retrogression [3]. Howeve , data mining and machine learning techniques have recently been introduced in response to the limitations of conventional statistical methods and the scarcity of medical data. An association rule approach based on a dynamic particle swarm optimizer was proposed by Mangat and Vig [4], and its classification accuracy is 82.18%. In order to estimate the 1-year postoperative survival of LCPs, Saber Iraji [5] examined the precision of adaptive fuzzy neural networks, extreme learning machines, and neural networks. The findings indicate that an extreme learning machine has the maximum sensitivity (90.05%) and specificity (81.57%), respectively. The boosted support vector machine (SVM) technique was employed by Tomczak et al. [6] to forecast the postoperative survival of LCPs. This technique combines the benefits of cost-sensitive SVM and ensemble learning, and the -mean can reach 65.73%. As can be observed from earlier studies, the majority of them overlook the potential negative effects of an unbalanced data distribution on classifier performance.

A class imbalance occurs when one class of data in a dataset is much greater than the others [7]. For balanced data, standard machine learning classifiers work well, but they fall short for unbalanced data. In particular, as medical technology has advanced, the number of long-term survivors following surgery for LCPs is significantly higher than the percentage of short-term deaths. For surviving (the majority class), this will result in better prediction accuracy, whereas for deaths (the minority class), it will result in worse recognition. Therefore, a strategy for predicting postoperative survival of LCPs must be proposed that has good classification performance for both surviving and deceased ones [2].

The two primary study methods in the present articles on unbalanced data processing approaches are data level and algorithm level [8].data-level is cleaning up or upgrading the dataset to increase the performance of the machine learning model.algorithm-level is adjusting the underlying machine learning algorithm or model to enhance performance for particular tasks.To enhance SMOTE, various methods address its limitation in ignoring data distribution and noise. SMOTE-RSB combines SMOTE with rough set theory to generate synthetic samples and remove noise [9]. SSMNFOS employs stochastic sensitivity measurement for noise filtering and oversampling. CURE-SMOTE uses clustering to eliminate noise. However, these often require parameter tuning [10]. CVCF-SMOTE, proposed in this paper, utilizes an ensemble-based filter called CVCF to identify and eliminate noise before applying SMOTE, reducing the risk of parameter-related errors [11].

Additionally, LC postoperative survival has not been effectively predicted using SVM, one of the most advanced classifiers. Due to its superior performance, SVM has been extensively used in earlier studies for statistical classification and regression analysis [12]. Some research combine resampling technology and SVM to handle imbalanced data in light of the constraints of SVM.

Particle swarm optimization (PSO) is used to automatically improve SVM (Support Vector Machine) parameters, resolving the issue of manually defining parameters in research. PSO-optimized SVM is used due to its ease of use and quick resolution. For instance, a Switching Delayed PSO-optimized SVM performs better than other SVM variations in the diagnosis of Alzheimer's disease [13]. However, choosing the best PSO parameters, such as particle size and inertial weight, can be difficult and time-consuming. If these parameters are chosen wrong, SVM performance may suffer. Improvements in PSO technology have produced new techniques as mulitvauled PSO-optimized SVM for feature selection, improving classification accuracy, and resolving convergence problems.

We suggest a two-stage hybrid technique to enhance the performance of the postoperative survival prediction of LCPs based on the enhanced SMOTE and FPSO-SVM. To enhance SMOTE's performance in the first step, noise samples are removed using CVCF.The postoperative survival of LCPs is predicted using FPSO-SVM in the second stage.

This study uses a two-stage methodology to improve postoperative survival prediction for patients with lung cancer. It uses SMOTE to balance datasets and CVCF and C4.5 to handle noisy data. For classification, Structural Risk Minimization theory-based Support Vector Machines (SVM) are used. By using Fuzzy Particle Swarm Optimisation (FPSO) to optimise the SVM's parameters, the accuracy of the model is increased. For reliable medical diagnosis, especially in circumstances of unbalanced datasets, this integrated strategy shows promise.

Existing method:

Data description:

The researchers [6],[2] utilized the thoracic surgery dataset from the Wroclaw Thoracic Surgery Centre, comprising 470 patient samples who underwent lung resection for primary lung cancer from 2007 to 2011. The dataset is imbalanced, with 400 patients surviving more than a year and 70 less than a year. Selected features from 36 preoperative predictors, chosen via the information gain method, are employed to predict whether patients' postoperative survival exceeds one year.

Data preprocessing:

SMOTE, a method to balance imbalanced data, can make data noise worse. To fix this, the researcher [2] suggested to use CVCF with C4.5 as the base classifier. CVCF spots noisy data used for multiple classifiers, and C4.5 handles data well which is suitable ensemble learning [14,15]. SMOTE adds synthetic samples to the minority class to balance data.

SMOTE generates synthetic samples for the minority class in imbalanced datasets. It selects the nearest neighbours for each minority instance, combines features with randomness, and adds these synthetic samples until the desired balance is reached, improving the dataset's suitability for machine learning. Support Vector Machine (SVM) is a powerful supervised classifier that thrives in high-dimensional data without overfitting. It finds a hyperplane in a higher-dimensional space to maximize the margin between different classes, aided by support vectors. [16,2]

The researcher [2] grounded in the theory of Structural Risk Minimization (SRM), emphasizing Support Vector Machines (SVM). SVM maximizes margins between data classes and employs kernel tricks for nonlinear data. The Lagrange formulation, parameter selection, and data transformation principles play pivotal roles in SVM's effectiveness.

The researcher [2] described the use of FPSO (Fuzzy Particle Swarm Optimization) to enhance SVM's classification performance by optimizing its parameters. Classification accuracy is employed as the fitness function, taking into account true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). FPSO adapts PSO dynamically using fuzzy logic, calculating particle parameters independently for better optimization. The method involves defining heuristic settings for particle swarms, updating particle positions and velocities, and fine-tuning parameters. The complexity of FPSO-SVM depends on the number of iterations, swarm size, and dimensionality of each particle in FPSO, as well as the SVM computations.

In this two-stage hybrid technique, LCP postoperative survival is predicted:Using Cross-Validation Clustering and Filtering (CVCF), chaotic data is first removed. The data is then balanced using SMOTE.In the second stage, LCP is predicted using the Feature Particle Swarm Optimization Support Vector Machine (FPSO-SVM). The dataset is divided up into subsets by CVCF, and C4.5 classifiers are trained using these subsets. Labels that are different from the rest of the samples are deleted as noise. In order to equalize the class distribution, SMOTE is applied to the cleaned dataset. The C and kernel hyperparameters of the FPSO-SVM are optimized via particle swarm optimization. The testing set for postoperative survival prediction is subjected to the upgraded SVM.By first cleaning the data and then adjusting SVM parameters,

this hybrid approach seeks to increase the precision of postoperative survival prediction in lung cancer patients (LCPs).

proposed methodology:

In this study, we are predicting the length of stay for 3 to 5 years for patients with lung cancer. Therefore, to modify our current methodology, we can acquire the dataset that includes the patient's medical history and a target variable that reflects the LOS for 3 to 5 years.After data collection, data preprocessing should be done by managing duplicates and missing values to clean up the data. data transformation to determine length of stay. Codify category variables, resize numerical features, and, if necessary, balance skewed data. Pick pertinent features, divide the data, and manage variables that depend on time. Normalise the target variable, analyse the data graphically, and continuously monitor the accuracy of the data.so instead of using a binary classification for regression problems, we should use a machine learning model. Regression algorithms like support vector, linear, and random forest regression can all be used.You'll utilise data to gauge your model's precision before applying it to forecast patients' hospital stays over the subsequent three to five years. Examine these predictions to identify the variables that affect longer or shorter stays. This clarifies why your predictions are what they are since you're dealing with fluctuating numbers rather than simple yes/no decisions.

Conclusion:

The hybrid methodology initially forecasts the survival of lung cancer patients for one year. Currently, we are working on applying the same process to forecast survival for three to five years. Additionally, parameter-free SVM models are being developed to enhance the accuracy of medical decision support over these timeframes.

Reference:

[1] Rotman, J. A., Plodkowski, A. J., Hayes, S. A., de Groot, P. M., Shepard, J. A. O., Munden, R. F., & Ginsberg, M. S. (2015). Postoperative complications after thoracic surgery for lung cancer. Clinical Imaging, 39(5), 735-749.

[2] Shen, J., Wu, J., Xu, M., Gan, D., An, B., & Liu, F. (2021). A hybrid method to predict postoperative survival of lung cancer using improved SMOTE and adaptive SVM. Computational and mathematical methods in medicine, 2021.

[3] Osuoha, C. A., Callahan, K. E., Ponce, C. P., & Pinheiro, P. S. (2018). Disparities in lung cancer survival and receipt of surgical treatment. Lung Cancer, 122, 54-59.

[4] Mangat, V., & Vig, R. (2014). Novel associative classifier based on dynamic adaptive PSO: Application to determining candidates for thoracic surgery. Expert Systems with Applications, 41(18), 8234-8244.

[5] Iraji, M. S. (2017). Prediction of post-operative survival expectancy in thoracic lung cancer surgery with soft computing. Journal of Applied Biomedicine, 15(2), 151-159.

[6] Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied soft computing, 14, 99-108.

[7] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications, 73, 220-239.

[8] Tsai, C. F., Lin, W. C., Hu, Y. H., & Yao, G. T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Information Sciences, 477, 47-54.

[9] Zhang, J., & Ng, W. W. (2018, October). Stochastic sensitivity measure-based noise filtering and oversampling method for imbalanced classification problems. In 2018 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 403-408). IEEE.

[10] Ma, L., & Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC bioinformatics, 18(1), 1-18.

[11] Afanasyev, D. O., & Fedorova, E. A. (2019). On the impact of outlier filtering on the electricity price forecasting accuracy. Applied Energy, 236, 196-210.

[12] Tao, Z., Huiling, L., Wenwen, W., & Xia, Y. (2019). GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. Applied soft computing, 75, 323-332.

[13] Zeng, N., Qiu, H., Wang, Z., Liu, W., Zhang, H., & Li, Y. (2018). A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease. Neurocomputing, 320, 195-202.

[14] S.-J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making," Journal of Biomedical Informatics, vol. 78, pp. 144–155, 2017.

[15] L. P. F. Garcia, J. Lehmann, A. C. P. L. F. de Carvalho, and A. C. Lorena, "New label noise injection methods for the evaluation of noise filters," Knowledge Based Systems, vol. 163, pp. 693–704, 2019.

[16] C. Cortes and V. N. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.