CLASSIFICATION OF LUNG CANCER USING SVM WITH FEATURE SELECTION BASED ON PSO-ROC

Abstract:

The global issue of lung cancer has grown to be very serious. Using machine learning to classify lung cancer is one method. The challenges in this study are how to apply Particle Swarm Optimization rate of change (PSO-ROC) as a feature selection method and support vector machine (SVM) as a classifier in the context of lung cancer classification; how to compare the accuracy values and running times between SVM without first reducing or selecting the features, SVM with PSO feature selection, and SVM with SVM with PSO-ROC feature selection in the context of lung cancer classification. The purpose of this work is to use SVM with feature selection based on the PSO-ROC algorithm to classify lung cancer. Three methods of classification were used in this study: first, Support Vector Machine (SVM) classification without feature reduction or feature selection; second, SVM and PSO feature selection method; and third, SVM and PSO -ROC feature selection. There are two categories for cancer: malignant and non-cancerous. The findings of this study should help the medical community categorize cancer more quickly and accurately, especially lung cancer. The PSO-ROC based feature selection selects limited number of attributes and yields high classification accuracy compare to others.

Introduction:

a) Lung cancer:

Approximately one in five malignancies in men and one in nine in women are lung cancers, making it the second most frequent type of cancer. Regretfully, although lung cancer incidence has been steadily declining in men over the past few years, it has been alarmingly climbing in women. Today, 42 out of 100,000 women have the condition, compared to just seven in 100,000 in 1940. And smoking is the reason, according to all the evidence. "How long it takes to get cancer depends on how many cigarettes you smoke every day," according to a field specialist. Nonetheless, research indicates that giving up smoking reduces the risk.

Lung cancer can be classified into two main types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), sometimes known as oat cell cancer due to the cells' resemblance to oat grains. The type of tumor that is identified determines the course of the disease and the available treatments. Given that the lungs are essential organs and that many forms of lung cancer grow and spread quickly, early detection and timely treatment—typically surgery to remove the tumor—are essential.

b) Types of Lung Cancer

Non-Small Cell Lung Cancer (NSCLC)

NSCLC is the classification given to the majority of lung malignancies. Squamous cell carcinomas make up around half of these (SCC). SCC, also known as epidermoid carcinoma, is more common in men and develops in the lining of the bronchi, which are the major airways. An other prevalent form of NSCLC is adenocarcinoma, which

develops in the lung's periphery. Large-cell carcinomas make up a tiny portion of NSCLC and typically originate in the smaller airways. Sometimes, non-small cell lung cancer that starts at the top of the lung spreads to the blood vessels and nerves that supply the arm.

NSCLC's three subtypes all develop in unique ways. The location and rate of spread of a given malignancy are major factors in treatment decisions.

- Approximately one-third of squamous cell or epidermoid carcinomas originate in the periphery of the lungs, although they typically occur in the bronchi in the core. Compared to other types of NSCLC, this one is more likely to result in bleeding and bronchial ulcers. The cancer cells usually divide every 180 days. While squamous cell carcinoma frequently invades adjacent tissue, it is less likely than other forms to spread quickly.

- While roughly 25% of adenocarcinomas originate near the lung's perimeter, the majority start in the center of the lung. These are little tumors, and every 180 days or so, the cells in them double. They are probably going to spread quickly. Bronchoalveolar adenocarcinoma is one type of the disease that starts in the alveoli and can spread to other sections of the lung through the airways.

- Massive tumors known as large cell carcinomas typically appear on the organ's periphery, though they can appear elsewhere in the lung. Throughout the course of the illness, the cells have the ability to penetrate the mediastinum and divide roughly every 100 days.

Small Cell Lung Cancer (SCLC)

Small cell lung cancer (SCLC) accounts for approximately 25% of lung malignancies. There are different variations of SCLC, such as oat cell cancer, which can consist of a combination of small cell and other cell types. These tumors have a rapid growth rate, doubling in cell number approximately every 30 days, and have a tendency to spread quickly to lymph nodes and other organs compared to non-small cell lung cancer.

c) PSO: Particle Swarm Optimization

PSO is best used to find the maximum or minimum of a function defined on a multidimensional vector space. Assume we have a function F(X) that produces a real value from a vector parameter (X, Y) (such as coordinate in a plane) and X can take on virtually any value in the space. The PSO algorithm will return the parameter which produces the minimum value of the function.

1. PSO is a stochastic optimization technique based on the movement and intelligence of swarms.
2. In PSO, the concept of social interaction is used for solving a problem.
3. It uses a number of particles (agents) that constitute a swarm moving around in the search space, looking for the best solution.
4. The figure 1 shows the flow diagram of PSO algorithms-

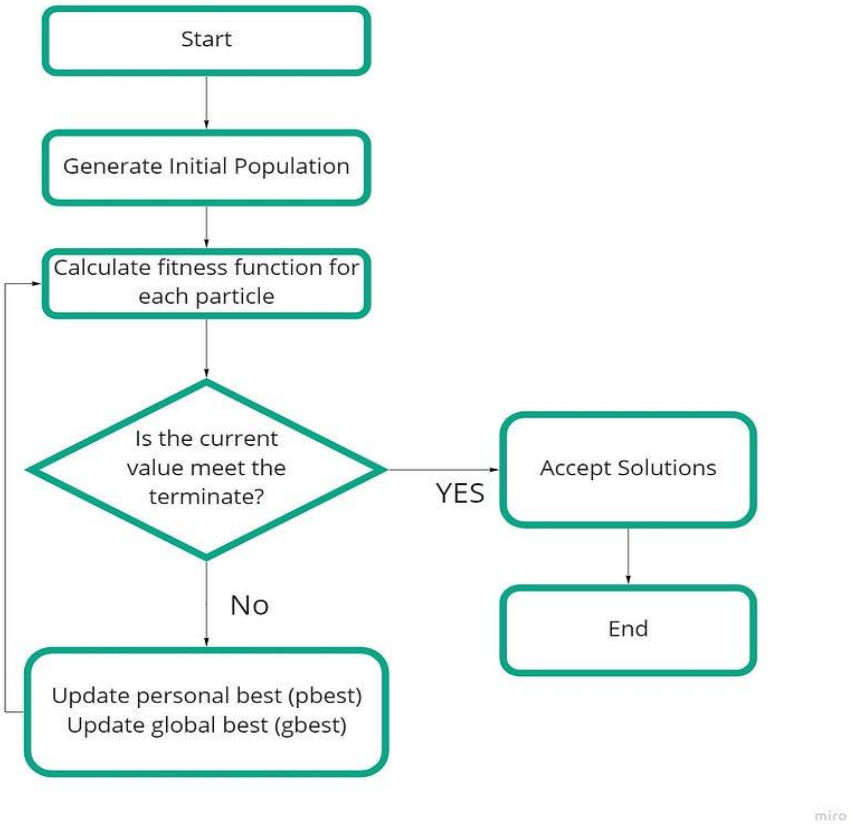

Fig 1. Flow diagram of PSO algorithms

d) Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

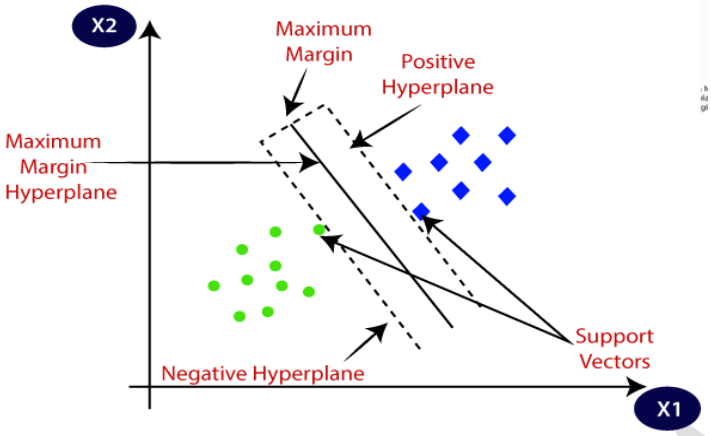

Fig 2. Support Vector Machine algorithm

Proposed Model: PSO-ROC with SVM:

The PSO-ROC based Feature Selection algorithm is implemented using the following steps:

1. Create a population of particles at first, each of which represents a subset of features.
2. Using the training and testing data, assess how well each particle's feature subset performs using an SVM classifier.
3. Using the performance evaluation as a basis, update each particle's personal best and global best positions.
4. Using the individual and global best positions, update each particle's velocity.
5. Continue repeating steps 2-4 until a stopping criterion is satisfied or for a predetermined number of iterations.
6. Decide which feature subset the global best particle best represents.
7. Utilizing the chosen features on the test data, assess the SVM binary classifier's performance.
8. To demonstrate the effect of feature selection, compare the outcomes with the SVM classifier that uses all of the features.

The figure 3 shows the graphical representation of the PSO-ROC based attribute selection with the help of SVM.

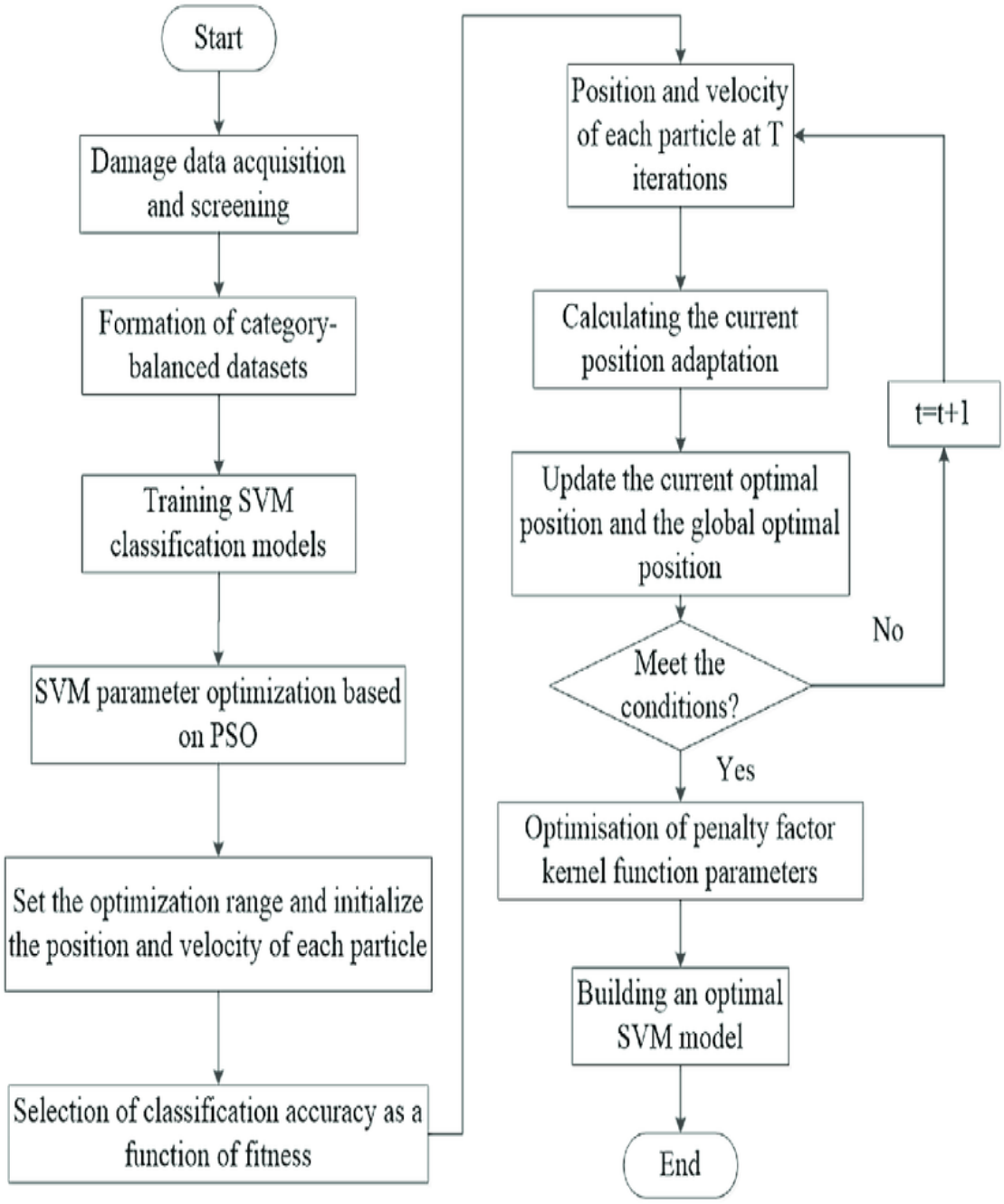

Fig 3. Feature selection (parameter optimization) process using SVM with PSO-ROC

Dataset:

Two datasets from the Kent Ridge Biomedical Dataset were utilized in this study. Specifically, the Michigan lung cancer data comprises 96 samples with 7129 features (genes), while the Ontario lung cancer data comprises 39 samples with 2880 features. The classification of cancer involves two classes: cancer and non-cancerous. It is hoped that the findings of this study will be beneficial to the community in achieving more precise and efficient classification of cancer, particularly lung cancer.

Result and Discussion:

The performance of the proposed model compared with the basic PSO feature selection model in terms of convergence iteration count and selected number of features with the best classification accuracy level. Table 1 shows the convergence iteration count and selected number of features using PSO and PSO-ROC with SVM. Table 2 shows the classification accuracy of Michigan and Ontario Cancer data sets with and without feature selection using SVM as the Classifier with PSO and PSO-ROC.

Table 1. Comparison of algorithm convergence and Selected Feature Set Count

Dataset	# Sample	# Features	PSO with SVM		PSO – ROC with SVM	
			Elected # features	Convergence Iteration count	Elected # features	Convergence Iteration count
Michigan lung cancer data	96	7129	48	≈145	26	≈57
Ontario lung cancer data	39	2880	32	≈82	18	≈34

Table 2. Comparison of classification accuracy

Dataset	# sample	# features	Classification Accuracy using SVM (%)	PSO with SVM (%)		PSO – ROC with SVM (%)	
				Elected # features	Accuracy (%)	Elected # features	Accuracy (%)
Michigan lung cancer data	96	7129	67.34	48	81.94	26	94.63
Ontario lung cancer data	39	2880	68.98	32	83.66	18	96.23

Table 2 shows PSO-ROC based feature Selection yields Superior classification accuracy result compare with PSO based feature selection model.

Conclusion:

This work investigates the use of Particle Swarm Optimization rate of change (PSO-ROC) as a feature selection technique and support vector machine (SVM) as a classifier for lung cancer classification. The objective is to compare the accuracy values and running times of three different approaches: SVM without feature reduction or selection, SVM with PSO feature selection, and SVM with SVM with PSO-ROC feature selection in the context of lung cancer classification. The goal of this work is to improve the way lung cancer is classified by utilizing SVM with feature selection based on the PSO-ROC method selects minimum number of important features with higher classification accuracy rate.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References:

[1] Z. Rustam, and S. A. A. Kharis, Journal of Physics: Conference Series 1442, 012027 (2020).

[2] Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer.

[3] American Cancer Society, Global Cancer Facts & Figures 3rd Edition (American Cancer Society, Atlanta, 2015).

[4] R. Chen, R. Manochakian, L. James, et al J Hematol Oncol 13, 58 (2020).

[5] M. Pulido, M. K. Derhartunian, Z. Qin, & E. M. Chung, Journal Of Neuroimmunology 299, 70–78 (2016).

[6] H. M. Alshamlan, G. H. Badr, & Y. A. Alohali, Computational Biology And Chemistry 56, 49–60 (2015).

[7] D. Karaboga, B. Basturk, Journal Of Global Optimization 39(3), 459–471 (2007).

[8] J. Luo, Q. Wang, & X. Xiao, Applied Mathematics and Computation 219(20), 10253–10262 (2013).

[9] M. Tuba, Latest Advances In Information Science And Applications, 252–257 (2012).

[10] S. Annuar, A. Selamat, & Z. Rustam, Journal of King Saud University 28(4), 395–406 (2018).

[11] S. Huang, N. Cai, P. P. Pacheco, et al., Cancer Genomics Proteomics 15(1), 41–51 (2018).

[12] WHO Classification of Tumours Editorial Board. Thoracic Tumours. 5th ed. Lyon, France: International Agency for Research on Cancer; 2021;(1).

[13] Nicholson AG, Tsao MS, Beasley MB, Borczuk AC, Brambilla E, Cooper WA, et al. Travis, The 2021 WHO Classification of lung tumors: Impact of advances since 2015. Journal of Thoracic Oncology. 2022;17(3):362-87.