CLASSIFICATION OF LUNG CANCER USING SUPPORT VECTOR MACHINE WITH FEATURE SELECTION BASED ON PARTICLE SWARM OPTIMIZATION

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.

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)

Most lung cancers are classified as non-small cell lung cancer (NSCLC). About half of these are squamous cell carcinomas (SCC). SCC, sometimes called epidermoid carcinoma, is more prevalent in men and arises in the lining of the large air passageways, or bronchi. Another common type of NSCLC is adenocarcinoma, which

occurs at the outer edges of the lung. A small percentage of NSCLC are large-cell carcinomas, which usually develop in the smaller bronchi. Non-small cell lung cancer that begins at the top of the lung sometimes spreads to the nerves and blood vessels leading to the arm.

All three subtypes of NSCLC develop differently. Treatment are often based on the location of the particular cancer and its rate of spread.

- Squamous cell or epidermoid carcinomas usually occur in the bronchi in the center of the lungs, but about a third of them arise on the periphery. This type of NSCLC is more likely to cause ulcers in the bronchi and bleeding than the other forms. Typically, the cancer cells double every 180 days. Although it often invades nearby tissue, squamous cell carcinoma is less likely to metastasize as soon as other types.
- Most adenocarcinomas begin in the middle of the lungs, but about 25 percent develop along the lung periphery. These tumors are small, and the cells double about every 180 days also. They are likely to metastasize early. The form known as bronchoalveolar adenocarcinoma develops in the alveoli and may spread through the airways to other parts of the lung.
- Large cell carcinomas are bulky tumors that usually develop on the organ's periphery; however, they can arise anywhere within the lung. The cells double about every 100 days and can invade the mediastinum during the course of the disease.

Small Cell Lung Cancer (SCLC)

About one in four malignancies involving the lungs are diagnosed as small cell lung cancer (SCLC). There are several types of SCLC or oat cell cancer, including a mix of small cell and other cell types. These cancers grow rapidly—doubling in cell number about every 30 days—and spread quickly to lymph nodes and other organs than the non-small cell type.

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. Below is a flow diagram of PSO algorithms-

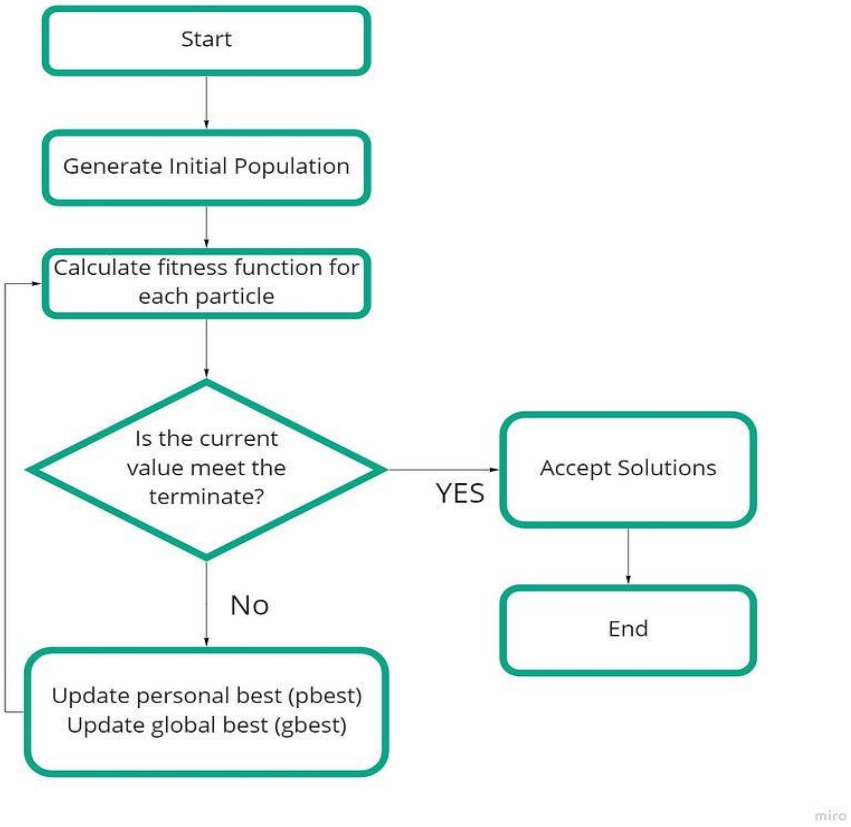

Fig 1. flow diagram of PSO algorithms

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:

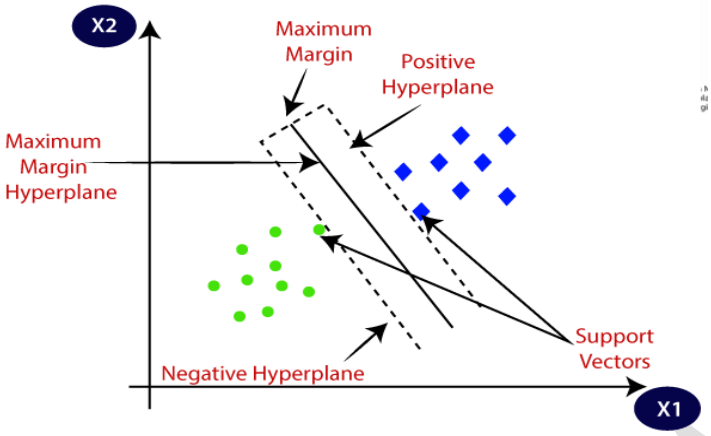

Fig 2. Support Vector Machine algorithm

PSO with SVM:

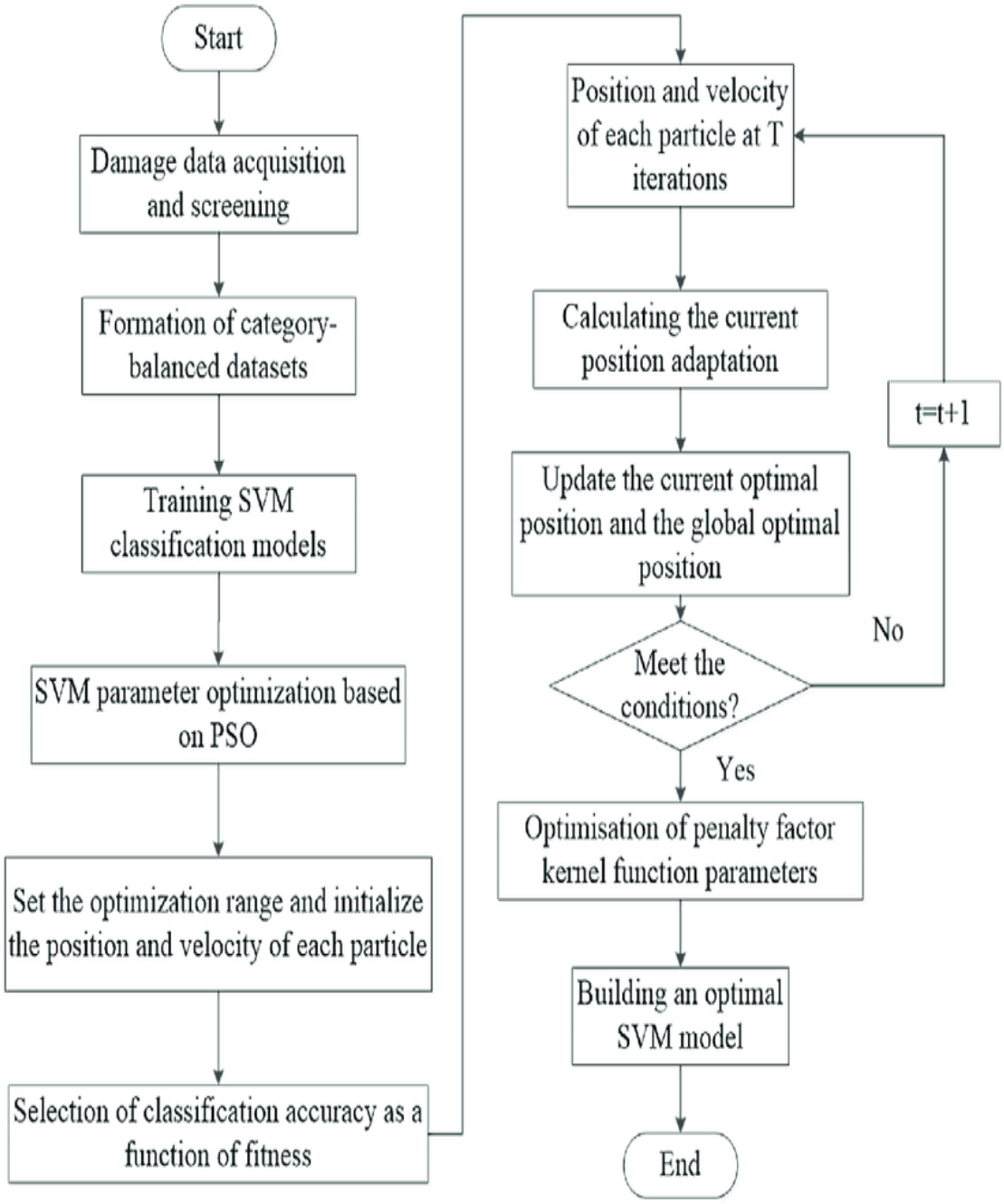

Fig 3. Data acquisition technique

Dataset:

This study uses two datasets from the Kent Ridge Biomedical Dataset, namely Michigan lung cancer data consisting of 96 samples with 7129 features (genes) and Ontario lung cancer data consisting of 39 samples with 2880 features. Cancer is classified into two classes, namely cancer and non-cancerous. Hopefully, the results of this study to be useful for the community in classifying cancer, especially lung cancer, more accurate and faster.

Table 1. Dataset of PSO – ROC with SVM (%)

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

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 in a superior way.

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).