Hybrid Systems (Soft Computing) and its use in classification of different types of Lung Cancer

Harjinder Singh
RegNo 11713154
Lovely Professional University, Phagwara, Punjab

Abstract

The most common cause of deaths (cancer-related), is Lung Cancer and also for both the genders male as well as females. As compared to other types of cancers like breast cancer, prostate cancer and pancreatic cancers, Lung cancer is responsible for more deaths. At the same time, it is also among the most curable cancer types if it can be diagnosed early. This paper presents a hybrid system with the help of soft computing, to classify lung cancer in different categories. We will introduce an intelligent system algorithm with help of PSO (Particle Swarm Optimisation) as well as Logistic Regression.
In feature extraction the particle swarm optimization plays a crucial role. Therefore PSO is proposed for selecting appropriate features for our classifier. Hyper parameters of Logistic Regression have an important role for recognising its accuracy however it is also one of best classifiers in manner of classifying complex data.

Keywords : Lung Cancer, PSO, Logistic Regression

Introduction

The 5-year survival rate for primary lung cancer is 19%, compared with 67%, 89% and 97% for colon, breast and prostate cancer respectively. More than one by two of the whole population with lung cancer will die within 1 year of diagnosis. This amount can be reduced if diagnosed at an early stage, 5-year survival reaches approximately to 50%.

Air pollution, Alcohol consumption as well as Smoking are some of major causes of Lung cancer in both genders male as well as female. Patients with Lung cancer are more seen to suffer with chest pain and dry cough.

Primary Lung cancer is classified in three levels or stages, These three classes or levels are as Low, Medium and High.

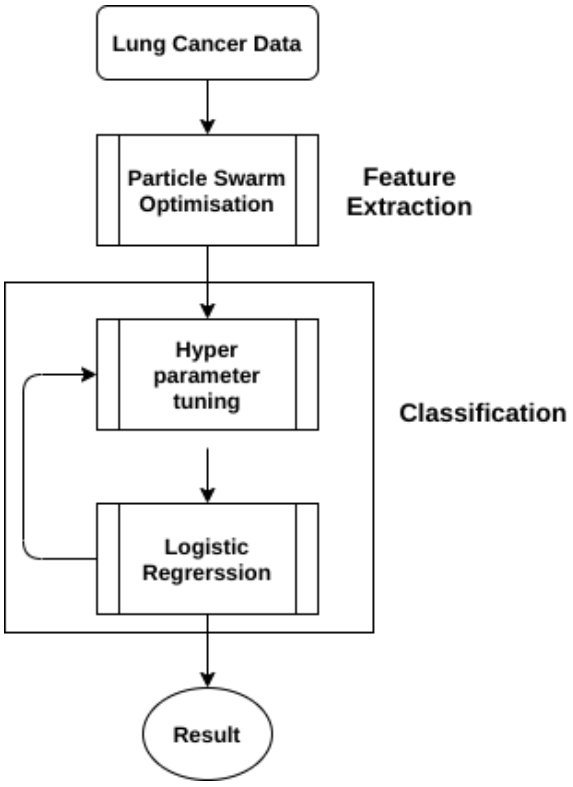
With the aim to classify levels of Lung Cancer we have come up with a hybrid algorithm using soft computing as well as logistic regression as our Hybrid System. Here we will use Logistic Regression for classifying different levels of Lung Cancer on the basis of different parameters as well as different features and alot of data.

Literature Review

Lung cancer in Australia is the leading cause of death due to cancer and is the second leading cause of all deaths in men and the fifth highest cause of all deaths amongst women. In 2006 there were 9563 cases and 7397 deaths due to lung cancer in Australia and only 12% of patients survived 5 years. The association between lung cancer (LC) and interstitial lung disease (ILD) can be explained by the shared risk factors like smoking and physiopathology of fibrogenesis and cancerogenesis. The relative LC risk is shown to be 3.5- to 7.3-times higher in ILD, with LC occurrence estimated at 10–20% in ILD, with >15% of ILD patients likely to die from LC. ILD incidence upon LC diagnosis varied from 2.4–10.9%. Primary radiological presentations consist of peripheral lesions, mostly in the inferior pulmonary lobes, either close to or within the ILD areas.

Proposed Methodology

A methodology proposed in this paper is illustrated below using a flow chart.



This methodology represents the execution of two processes , Feature extraction and Classification.

We will look at both of the processes that we have mentioned above. As well as with help of some data and graphs we will make our algorithm more useful as compared to classic algorithms. As a Hybrid algorithm our system possesses more accuracy as compared to classic classifiers.

Feature Extraction

For feature extraction we will be using Particle Swarm Optimization (PSO) algorithm.

The PSO algorithms is given as :

Step1: Cluster and Iterations are initializing

Step2: Parameter ***P***, ***sc***, ***fc***, numsucc=0, and numfail=0 are initializing

Step3: Identify a fitness function

Step4: To find the fitness of each particle rate

Step5: Update the local best solution.

Step6: Steps 4 and 5 repeat

Step7: Each particle velocity and position are updated

Step8: Execute the selection operator

Step9: If any local best position y_i has changed, and to perform the clustering algorithm.

Step10: End procedure.

The above algorithm of Particle Swarm Optimisation can be used to extract features as it updates the values of the equation.

In this method focuses on the current best position of a new particle, the new particle is considered as the swarm and the velocity update equation for new particle is defined as:

$$v\varphi(t+1) = x\varphi(t) + pbest(t) + \omega v\varphi(t) + p(t)(1-2r)$$

The gbest position is to improve the random search area around the position. The **r** and **p(t)** is a random vector and diameter of the search area. The range of the random vector lies between **0** and **1**. The diameter of the search area can be updated using the following equation:

$$\rho(t+1) = \begin{cases} 2\rho(t), & \#successes > sc, \\ \left(\frac{1}{1.5}\right)\rho(t), & \#failures > fc, \\ \rho(t), & otherwise, \end{cases}$$

Where,

sc and **fc** - Threshold parameters, the value of sc=15 and fc=15

Classification

For classification purposes we are using the logistic regression. Given a data(X,Y), X being a matrix of values with m examples and n features and Y being a vector with m examples. The objective is to train the model to predict which class the future values belong to. Primarily, we create a weight matrix with random initialization. Then we multiply it by features.

Logistic regression calculates the probability of a particular set of data points belonging to either of those classes' given the value of *x* and *w*. The logic is that say, we have a set of values that we obtain from negative infinity to positive infinity based on the linear model, we need to narrow it down to a score that is in between zero and one as probabilities always are in that range and logistic regression talks about probabilities. The link function, sigmoid function takes care of this work.

The use of exponent in the sigmoid function is justified as probability is always greater than zero and the property of exponents takes care of this aspect. Then we need to worry about limiting the values less than one, which is done by dividing the value in the numerator by value greater than it.

If we take into consideration the conditional probability of getting an output P(y=1|x;w) is equal to the sigmoid function and the p(y=0|x;w) = 1-p(y=0|x;w) and if take in that our sample has a Bernoulli distribution then the cost function for the logistic regression model is derived by,

$$p(y|X;W) = \sum_{i=1}^{i=n} (h_W(X))^y + (1 - (h_W(X))^{1-y})$$

Taking log this equation can be transformed into the cost function above. The presence of the minus sign in the beginning of the function is to ensure we try minimizing the negative of likelihood instead of maximizing the value as gradient descent minimizes the error.

Result and Discussion

In order to display fidelity and capacity of the suggested algorithm, PCO + Logistic Regression, a comparison with other algorithms has taken place, concerning highest correctness rates, which are demonstrated. In the first method, microarray data has been classified directly with the SVM method. In the second method, all PCO selected features have been employed to train Logistic Regression. As can be seen, the proposed algorithm yields the highest value of correctness rate in comparison with other methods in

datasets (Lung cancer datasets). As an illustration, our proposed algorithm exhibits relative improvements of 3.3% over PCO + Logistic Regression and Logistic Regression algorithms in the Lung cancer dataset. Furthermore, it is obvious that if all ICs are used to reconstruct new samples, the correctness rate of the sub-classifier will not always be better than employing Logistic Regression directly, while, with selecting an appropriate set of features using PCO, the result improves.

Conclusion

In this paper, different segmentation algorithms have been classified to early detection of lung tumors. Analyzing these techniques, which include feature extraction using Particle Swarm Intelligence, the better accuracy result of the tumor detection is enhanced proposed method with maximum accuracy rate of 95%. The proposed method is more accurate when compared to the existing segmentation algorithm. In future studies, various segmentation algorithms will be integrated and to improve the better accuracy rate.

References

1. Lavanya M, Muthu Kannan P "Lung cancer segmentation and diagnosis of lung cancer staging using MEM (modified expectation maximization) algorithm and artificial neural network fuzzy inference system (ANFIS)" ISSN 0970-938X Biomed Research 2018 29 (14): 2919-2924.

2. Dehnavi AM, Sehhati MR, Rabbani H. Hybrid method for prediction of metastasis in breast cancer patients using gene expression signals. J Med Signals Sens

3. Carpentier AS, Riva A, Tisseur P, Didier G, Hénaut A. The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. Comput Biol Chem.

4. Moffy Crispin Vas, Prof. Amita Dessai "Classification of Cancerous and Non-Cancerous Lung Cancer Nodules Using Image Processing Techniques" International Journal of Advance Research in Science and Engineering Vol. 6 Issue 4 April 2017

5. Hamiton HJ, Shan N, Cercone N. RIAC: A rule induction algorithm based on approximate classification. In International conference on engineering applications of neural networks, University of Regina

6. Neelam Marshkole, Bikesh Kumar Singh,and Thoke, A.S., “Texture and Shape based Classification of Brain Tumors using Linear Vector Quantization” International Journal of Computer Applications,Vol.30