

Novel Genetic Algorithm for Early Prediction and Detection of Lung Cancer

Ammar Odeh¹, Ibrahim Al Atoum¹, Abraham Bustanji²

¹Department of Computer Science and Information Systems, Al Maarefa Colleges for Science & Technology, Riyadh, Kingdom of Saudi Arabia

²Department of Medicine, College of Medicine, Almaarefa Colleges for Science and Technology, Riyadh, Saudi Arabia

Email address:

aoudah@mcst.edu.sa (A. Odeh), iotoum@mcst.edu.sa (I. A. Atoum), ibustanji@mcst.edu.sa (A. Bustanji)

To cite this article:

Ammar Odeh, Ieee Member, Ibrahim Al Atoum, Abraham Bustanji. Novel Genetic Algorithm for Early Prediction and Detection of Lung Cancer. *Journal of Cancer Treatment and Research*. Vol. 5, No. 2, 2017, pp. 15-18. doi: 10.11648/j.jctr.20170502.13

Received: October 29, 2016; **Accepted:** March 14, 2017; **Published:** March 22, 2017

Abstract: Nowadays, many researchers try to find out a system that enables to detect and expect diseases early so as to find the appropriate precaution or medical treatment of it. One of the leading causes of death worldwide is Cancer. Most of the deaths from this disease are due to late prediction and detection. According to the American Cancer Society (ACS); lung cancer is the second most common cancer; it accounts for about 13% of all new cancers. It is expected to have a 221, 200 new cases of lung cancer in 2015 with 158, 040 estimated deaths from lung Cancer [1]. The main objective of this study is to reach the highest accuracy and speed of its predecessors and this is what has been obtained.

Keywords: Cancer, Crossover, Mutation, Genetic Algorithm, WEKA, OWL

1. Introduction

Effective detection of lung cancer in its early stages is considered as a significant aspect for rescuing the lives of many patients by enabling patient to avoid risk factors. In this paper a machine learning technique is applied for early prediction of lung cancer and detection. Machine Learning (ML) is a branch of artificial intelligence (AI) that uses a variety of statistical, probabilistic and optimization methods which enables the computer to learn from past examples. One of these optimization methods is Genetic Algorithms (GA) that uses the combination of selection, recombination and mutation to evolve a solution to a problem. This research uses this novel technique to generate early solutions for lung cancer disease. Due to non-deterministic of genetic algorithm we enable to deal with the gene variation with highly accuracy and minimize searching time. In [2] a GA is used as a method of feature (genes) selection for the support vector machine and artificial neural network to classify lung cancer status of a patient. In [3] a system is developed to test the cancer risk level using data mining. In [4] three classifiers namely Support Vector Machine (SVM), Artificial Neural Network (ANN), and K-Nearest Neighbour (k-NN) are

applied for the detection of lung cancer to find the severity of disease (stage I or stage II).

This paper is organized as follows. Section II, presents the related work of existing lung cancer detection algorithms. In section III, provides the proposed algorithm. In section IV, offers analysis of the proposed algorithm. Finally, section V, offers findings and discussions.

2. Related Work

Authors in [5] demonstrated the evaluation method for identifying predictive Gene. They merged between the GA combination and KNN. Then they rank the GA output. Based on ranking process, system applied Leave One-out Cross Validation (LOOCV) for error estimation. The main advantage of this model has a good accuracy on training dataset.

In [6] provides the modelling of the lung cancer Ontology in OWL 2. The proposed system represents online clinical decision support application. The proposed algorithm classifying patients' situation then producing the guide line

based treatment recommendations with the help of ontological inference. The key objective of Lung Cancer Assistance LCA assists clinicians by inferring and analysing existing patient data to make a meaningful outcome based on clinical guideline rules.

A new algorithm in [7] is introduced by using different MATLAB procedures to implement the different classes of Image processing. Initially a pre-processing technique is applied on the captured Image to improve image quality. Secondly, a segmentation method is applied partition the under consideration image into multiple segments. Finally, a Feature extraction process is implemented to get more accurate results by using various enhancement and segmentation techniques.

A novel algorithm presented in [8] by employed two segmentation processes, Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm, for segmenting sputum color images to detect the lung cancer in its early stages. Based on research conclusion the HNN segmentation results are more accurate and reliable than FCM clustering in all cases. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will increase the probabilities of survival for the patient.

Start

Input dataset [1...N]

Do

Selection Step: Rank with S2N
Split data into rows and columns

Compute Mean (M) for each vector
$$\mu = \frac{\sum x_i}{n}$$

Compute σ for each vector
$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Compute S2N for vector in each group
Select the best k feature

Crossover Step:

Single point crossover to exchanging the data between parents to make new generation of genes

Mutation Step

Twos mutation is a tactic, where two genes are chosen randomly and their positions in the chromosome are exchanged.

Until we find the optimal accuracy, which will be near 100%.

End

In [9] a new proposed system presented, the system aims to early detecting the lung cancer. This system can achieve the pre-processing process based on CT and BET captured images. Additionally it can improve the image quality by using histogram equalization and Median filter. Thereafter a Feature selection and extraction process is accomplished and followed by analysing the output data to classify the features.

Finally the system has the ability to suggest therapy options.

3. Proposed Algorithm

Genetic algorithm provides the optimal solution based on fitness function. Our proposed algorithm includes feeding the GA with an initial selection of population. These populations come from raw dataset of Microarray which represents the components of chromosome. The next step is calculating the fitness function. Based on output of these calculations; the higher fitness could be reserved and discarding the lower one. Third step is the crossover process, that responsible to yield good generation by combine the best component from different genetic. Finally, mutation process is implemented which will generate new gene structure with small random probability. Figure 1 illustrates the proposed system components.

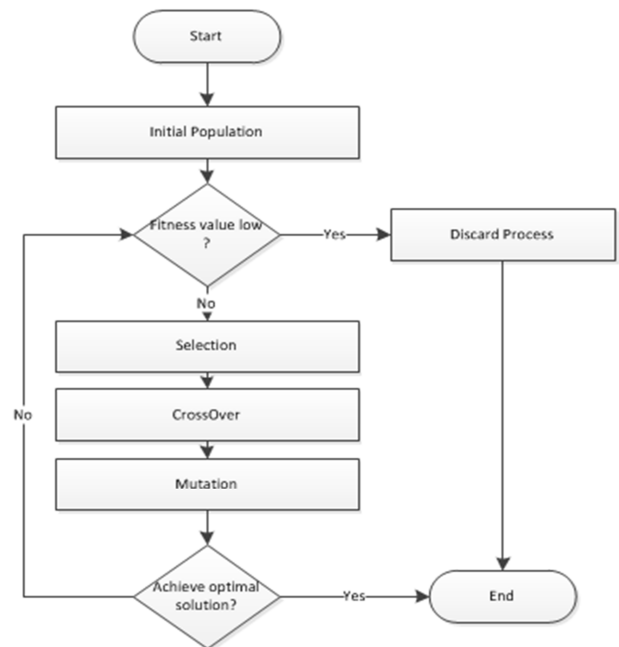


Figure 1. Proposed Genetic Algorithm Components.

The previous flowchart shows the general form of the algorithm that will be followed in this study and each stage of this flowchart have own technique. The first is to follow the flowchart so as to get satisfactory results regardless of the used technique. The below steps clarifies the different stages of the proposed GA.

4. Analysis of the Proposed Algorithm

In our proposed algorithm we used a training data provided by Michigan, Harvard, and Ontario universities and by using Waikato Environment for Knowledge Analysis (WEKA) [10]. The data consist of rows and columns, based on Mean and Standard deviation calculation we can find out Signal-to-Noise (S2N).

Signal-to-Noise ratio enables system to find out the best features determines lung cancer probability. The proposed

system then apply single point crossover to generate different offspring. Figure 2 exemplifies a single crossover.

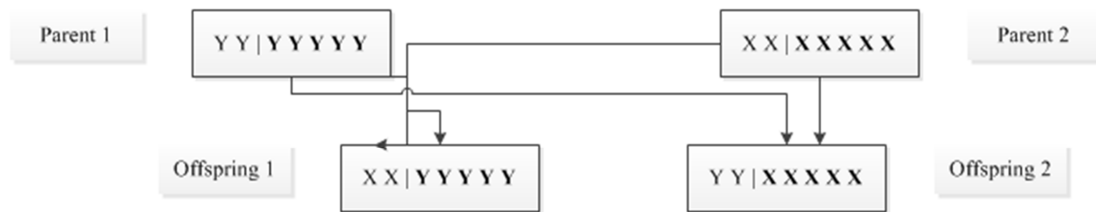


Figure 2. Single point crossover.

The next step is Twors mutation is a tactic, where two genes are chosen randomly and their positions in the chromosome are exchanged. Figure 3 illustrates the Twors mutation process.



Figure 3. Mutation operator TWORS.

The reevaluation process for the new generations enables us to estimate the best choice and figure out the optimal gene that indicate the lung cancer disease.

5. Result and Discussion

Based on our simulation results the proposed system needs 22.23 second to be built by using WEKA software. Table 1 compares between the required times needed to build a system in four other categories. In the proposed algorithm we need 22.23 seconds which represents the best second value comparing to the remaining algorithms.

Table 1. Cancer predicate Algorithms and processing time.

Algorithm	Process Time / second
Michigan [11]	5.85307
Harvard [12]	33.90013
Harvard2 [12]	24.41596
Ontario [13]	40.81986
Proposed algorithm /Ontario	22.23

Moreover the selection process is used to create a two sets of high probability and the lowest one it's just consume 0.91 second as shown in Figure 4.

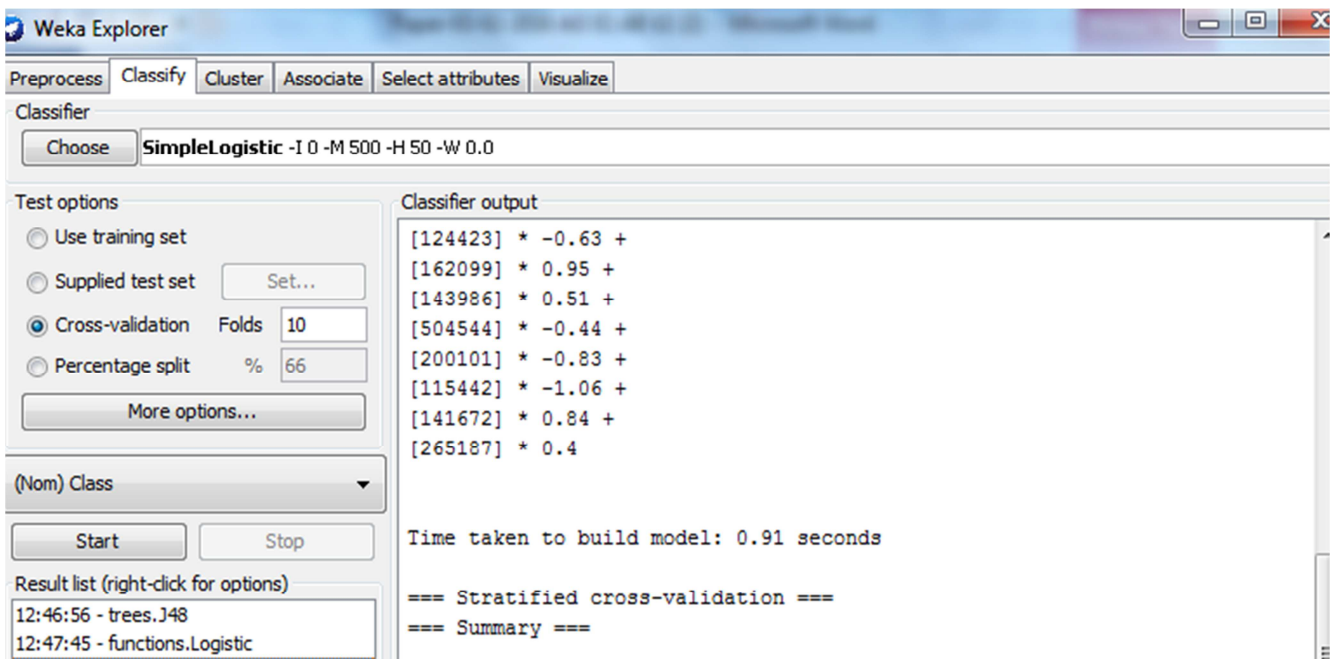


Figure 4. Selection processing time.

Based on simulation output after 16 gene generations the proposed system achieves 84% of accuracy where the maximum presented accuracy is 81. Figure 5 exemplifies the number of generation and accuracy percentage.

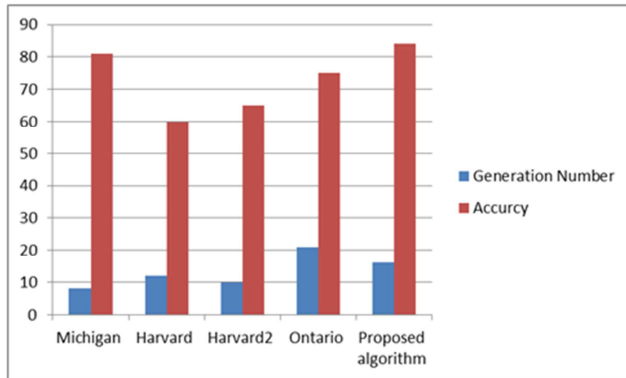


Figure 5. Number of Generation and Accuracy.

6. Conclusion

Lung cancer is one of the most risky diseases in the world. Accurate Diagnosis and early detection of lung cancer can raise the survival rate. Our novel system includes feeding the GA with an initial selection of population and then selects the best choice to generate the new springer by using crossover operation. Then some mutation enables system to do some modification on offspring. Our results show high accuracy ratio 84% comparing to other algorithms, in addition the proposed system did not consume along processing time 22.23 seconds.

Our future plans we can optimize of the classification process of tissue samples by using filling missing gene data list. Based on optimization process we can find out a new genetic algorithm that enables us to early detection and higher accuracy percentage.

References

- [1] M. Boeckh, W. Leisenring, S. R. Riddell, R. A. Bowden, M.-L. Huang, D. Myerson, T. Stevens-Ayers, M. E. Flowers, T. Cunningham, and L. Corey, "Late cytomegalovirus disease and mortality in recipients of allogeneic hematopoietic stem cell transplants: importance of viral load and T-cell immunity," *Blood*, vol. 101, pp. 407-414, 2003.
- [2] F. Taher and R. Sammouda, "Lung cancer detection by using artificial neural network and fuzzy clustering methods," in *GCC Conference and Exhibition (GCC)*, 2011 IEEE, 2011, pp. 295-298.
- [3] M. V. A. Gajdhane and L. Deshpande, "Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques."
- [4] O. Abdoun, J. Abouchabaka, and C. Tajani, "Analyzing the Performance of Mutation Operators to Solve the Travelling Salesman Problem," *arXiv preprint arXiv:1203.3099*, 2012.
- [5] S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, and L. Chen, "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines," *FEBS letters*, vol. 555, pp. 358-362, 2003.
- [6] W. F. Baile, R. Buckman, R. Lenzi, G. Glober, E. A. Beale, and A. P. Kudelka, "SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer," *The oncologist*, vol. 5, pp. 302-311, 2000.
- [7] M. S. AL-TARAWNEH, "Lung Cancer Detection Using Image Processing Techniques," *Leonardo Electronic Journal of Practices and Technologies*, vol. 11, pp. 147-58, 2012.
- [8] J. Schneider, G. Peltri, N. Bitterlich, K. Neu, H.-G. Velcovsky, H. Morr, N. Katz, and E. Eigenbrodt, "Fuzzy logic-based tumor marker profiles including a new marker tumor M2-PK improved sensitivity to the detection of progression in lung cancer patients," *Anticancer research*, vol. 23, pp. 899-906, 2002.
- [9] J. Schneider, N. Bitterlich, H.-G. Velcovsky, H. Morr, N. Katz, and E. Eigenbrodt, "Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer," *International journal of clinical oncology*, vol. 7, pp. 145-151, 2002.
- [10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, pp. 10-18, 2009.
- [11] A. Giatromanolaki, M. Koukourakis, E. Sivridis, H. Turley, K. Talks, F. Pezzella, K. Gatter, and A. Harris, "Relation of hypoxia inducible factor 1 α and 2 α in operable non-small cell lung cancer to angiogenic/molecular profile of tumours and survival," *British journal of cancer*, vol. 85, p. 881, 2001.
- [12] H. Ji, M. R. Ramsey, D. N. Hayes, C. Fan, K. McNamara, P. Kozlowski, C. Torrice, M. C. Wu, T. Shimamura, and S. A. Perera, "LKB1 modulates lung cancer differentiation and metastasis," *Nature*, vol. 448, pp. 807-810, 2007.
- [13] F. A. Shepherd, J. Dancey, R. Ramlau, K. Mattson, R. Gralla, M. O'Rourke, N. Levitan, L. Gressot, M. Vincent, and R. Burkes, "Prospective randomized trial of docetaxel versus best supportive care in patients with non-small-cell lung cancer previously treated with platinum-based chemotherapy," *Journal of Clinical Oncology*, vol. 18, pp. 2095-2103, 2000.