Prediction of Lung Cancer using Data Mining Techniques

F. Leena Vinmalar[1]
Ph.d Scholar, Department of Computer Science
Chikkanna Govt Arts College,
Tirupur.

Dr. A. Kumar Kombaiya[2]
Assistant Professor, Department Of Computer Science
Chikkanna Govt Arts College,
Tirupur.

***Abstract*:- Cancer is very dangerous and common disease that causes death worldwide. Early diagnosis of cancer provide more possibility of getting cured. Cancer disease generates abnormal growth of cells which spreads to all parts of body. In this paper we discuss, the early prediction of lung cancer with help of data mining techniques. Lung are spongy organs that affected by cancer cells that leads to loss of life. The common reasons of lung cancer are smoking habits, working in smoke environment or breathing of industrial pollutions, air pollutions and genetic. In this paper we have proposed a genetic algorithm based dataset classification for prediction of multiple models. The usage of genetic algorithm (GA) have shown better performance when compared with Particle swarm optimization and differential evolutions.**

Keywords- Data mining, Classification, Weighted Average, Genetic Algorithm.

I. INTRODUCTION

Early prediction of lung cancer helps to prevent the lung to be affected more by the dangerous cells. Any cancer identified in the early stages is very significant to provide better treatment and gives a huge positive progress to get ride from cancer. Especially the lung cancer are caused due to smoking and pollution, so early detection can help the patients to stop from smoking or from other factor that causes the cancer. Data mining is a part of Artificial Intelligence that uses a variety of data sets, probabilistic and mining models which provides a technique to predictive results using past results. One of the very famous and optimized method in the data mining is Genetic Algorithms (GA) that uses the collections of selections, recombination and a model to evolve a solution to a problem. In this paper we have used genetic algorithm to early predict the lung cancer diseases.

In recent years, lung cancer prediction done by using methods like ensemble classification approaches by using Decision Tree, Naive Bayes, and Classification based on Multiple Association Rule (CMAR) method on the basis of the voting method [16]. A method like the ensemble method of Multilayer Perceptron, Random Forest and Random Tree are used for lung cancer prediction [10]. Support vector machine
(SVM), Naive Bayes, KNN and C4.5 method used for lung cancer prediction out of which SVM outperforms [4] . Principle component analysis based ANN hybrid model is used for lung cancer prediction [9]. A method like interpretable models [6] used for lung cancer prediction. SVM and K-nearest neighbour approach proposed for lung cancer prediction [8]. A method like image processing in the first phase and Back Propagation Neural-Network and logistic regression method used for lung cancer prediction [2]. Bayesian Network and SVM used for lung cancer prediction carried out using Weka tool [3]. Deep learning SVM
(D-SVM) approach is used for lung cancer prediction [19]. K-means clustering and decision tree method used for lung cancer prediction [7]. Method like Random Forest (RF), Radial Basis Function Network
(RBF) and Neural Network (NN) is used as a base learner and for ensemble AdaBoostM1, Real Ada Boost, and Multi Boost AB were used [1]. Ensemble method using the random forest for lung cancer prediction [11].

C4.5 Decision SVM and Naive Bayes with effective feature selection techniques used for lung cancer prediction [15]. A method like Random Forest and Naive Bayes gives better result in lung cancer prediction [20]. Decision tree used in lung cancer prediction [18]. For increasing the accuracy of the model ensemble method is to be used. Preferably assigning weights manually is not a good strategy. So, in recent years, iDHS-EL and iDNA-KACC-EI methods with ensemble approach are used to a find weight for the fusion process [12]. iRSpot-EL method of ensemble used for fusion calculates the distance by utilizing the affinity propagation algorithm [13].

II. DATASETS AND MODEL

In this work we have used the dataset from UC Irvine Machine Learning repository, the sample data set and its attributes are shown below

Dataset and Description Table 1.

Data ID	Data Identification No.
Age	(Age between 13 – 80)
Gender	(Male -1/ Female -2/Others-3)
Air Pollution	(Range 1 -10)
Alcohol use	(Range 1 -10)
Occupational	(Range 1 -10)
Hazards	(Range 1 -10)
Genetic Risk	(Range 1 -10)
chronic Lung Disease	(Range 1 -10)
Balanced Diet	(Range 1 -10)
Obesity	(Range 1 -10)
Smoking	(Range 1 -10)
Passive Smoker	(Range 1 -10)
Chest Pain	(Range 1 -10)
Coughing of Blood	(Range 1 -10)
Weight Loss	(Range 1 -10)
Snoring	(Range 1 -10)

III. METHODOLOGY

The proposed methodology is represented in the Fig. 1. It is divided into three steps

1) Data Randomization and partitioned
2) Choosing the model, training and testing the model
3) Generate the weighted average of method and testing.

Best models selection is based on accuracy estimation. Top three model is used for the final ensemble method. Proposed approach methodology is represented in fig 1.

In final phase, we have to calculate weight by evolutionary algorithm named as genetic algorithm (GA) and optimize weight by GA applies to a weighted average method of an ensemble. We have taken three natures inspired algorithm (NIA) for the weight optimization out of which GA outperforms because it gives the best fitness function, best chromosome and function evolution is also less in comparison to both algorithm and time.

S.NO	MODEL	DESCRIPTION MODEL
1	SVM	A wrapper class utilized in classification and out linear detection
2	Decision Tree	An extension of C4.5 classification
3	Random Forest	Forest of Random decision trees
4	Linear Model	Statistical method used to create a linear model
5	SVM Poly	Similar to SVM model but here in place of radial polynomial method used.
6	Naïve Bayes	It is simple "probabilistic classifiers" based on Bayes theorem.

Table 2. Classification Models Descriptions

taken in weight optimization is also less so, it is used in weighted average ensemble method. This process is used to increase the efficiency of the proposed model. In classical weighted average method we have assigned weight manually. So, the accuracy of the model was not increasing. In our proposed solution we have calculated weight by GA. So, the accuracy of the model should be increased to comparing to classical method. Here we compare three NIA algorithm named PSO, DE, and GA. We use prediction of the top three model here with optimizing weight in weighted average ensemble method.

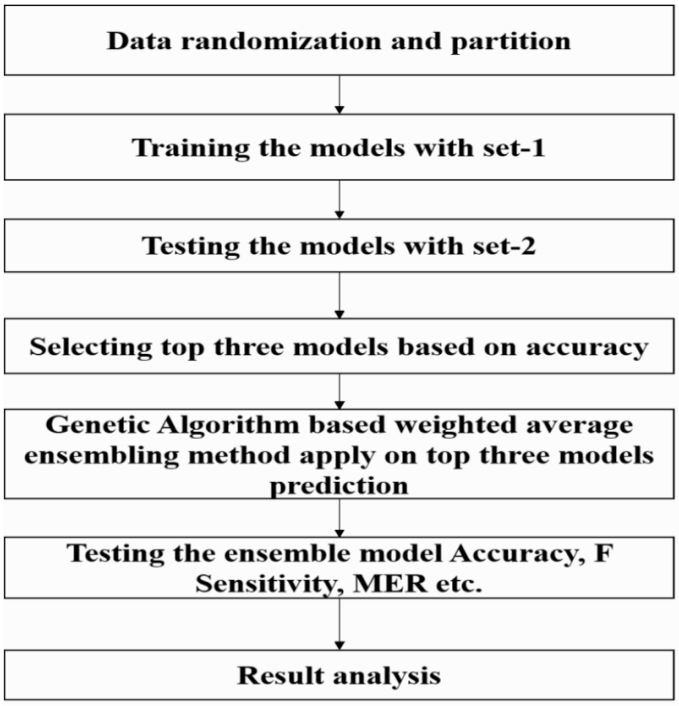

Fig.1 Proposed Methodology

A. Genetic Algorithm

An evolutionary algorithm (GA) is proposed to solve the problem of optimization of weight. GA is a meta heuristic algorithm which is inspired by nature. GA is used to generate high-quality solutions, for optimization and search problems which depend on bio inspired promoter named as mutation, crossover, and selection [5]. GA was instigated by J. HOLLAND who invented it in the early 1970's [5]. its based on natural evolution theory of Darwin. GA has been applied to several areas such as telecommunication, routing, scheduling.

It proves that the solution obtains by it are very effective [5].

GA algorithm as shown in 1.

Algorithm 1:

Genetic Algorithms Framework
begin n=0
Random initialization of population p(n)
The fitness of population determine p(t)
while n=n+1 do
Selection of parent from population p(n)

Crossover operation perform on parents to create off springs (n+1) Mutation operation perform (n+1)

The fitness of population determine
(n+1) end
while

In the genetic algorithm, the selection process is done to find the best fit of solution and eliminates the duplicate of data that play important role in selection of good procedure. After selection process the crossover is done to verify the accuracy

of the data. The crossover process produces multiple copies of good solutions that generate off-spring.

however, the possibility of generating better off-springs is more, since the off-springs have been made from those individuals which have been survived throughout the selection phase.

And finally Mutation having an extremely low probability takes part of an individual. If any bit of an individual is selected to be mutated then it is overturned with a feasible replacement value for that bit.

For keeping the diversity of the population mutation applied next to crossover phase. Mutation is a random change in off-spring. It is never giving better offspring always, how- ever it searches one or two solutions to the neighbourhood of the native solution.
We conclude that GA is a nature inspired based algorithm to solve optimization problems. GA constantly updates a population of individual solutions. At each phase, On the basis of fitness GA choose

individuals from the present population and generate a new off-springs by utilizing them as a parent for the upcoming generation.

IV. IMPLEMENTATION

In the proposed implementation, a random data is generated and partitioned. These dataset is process to generate a training model, the training model is tested for accuracy and a GA Based weighted average is obtained as result.

Training result of the experimented classification models and their performance parameter values as shown in table 3. Top three best model is selected for further processing of weighted average ensemble method on the basis of accuracy.

These top best three models are much accurate than other models which have a minimum error rate and high accuracy. Adaboost, SVM and random forest model are used for the ensemble. Predictions of these models used as a training data for the weighted average ensemble approach. Out of which GA outperforms due to which it is used for optimization of weight.

Here minimization of fitness function is done by GA. It gives best fitness 6.36 and best chromosome which optimizes weight is (.31, .21, .48). The parameters for PSO, DE, and GA are decided on the basis of the previous basic paper. The training of the ensemble approach uses the best top three models prediction as a training data. Top three models prediction we apply to weighted average method.

	Our GA approach	Decision trees (IG, Red-Err)	Decision trees with boosting (IG, Red-Err, 3 trees	Neural networks (1-of-N, batch learning)	Naive Bayes
Run 1	95.35	95.35	95.35	93.02	93.02
Run 2	97.67	93.02	97.67	97.67	90.7
Run 3	97.67	97.67	97.67	97.67	93.02
Run 4	95.35	95.35	97.67	97.67	88.37
Run 5	95.35	95.35	95.35	93.02	88.37
Run 6	97.67	97.67	97.67	97.67	88.37
Run 7	95.35	93.02	93.02	95.35	93.02
Run 8	97.67	95.35	95.35	95.35	90.7
Run 9	100	100	97.67	95.35	90.7
Run 10	95.83	93.75	97.92	95.03	85.42
Average	96.79	95.65	96.54	95.86	90.17
Standard deviation	1.59	2.23	1.67	1.46	2.53

Table 3 . show top three best models

V. CONCLUSION

In this study, a new method GA is introduced to excel the limitation of classical weighted average ensemble method. For this method, we experimented eight machine learning models and out of which select top three model based on accuracy for further process. To obtain maximum accuracy we proposed separate data partition. For ensemble purpose, we used here GA based weighted average method. The result obtained from experiments has been analysed across, data tables and graph plots. The result obtained proves that excelling the limitation of the classical weighted average method in terms of accuracy and other performance parameters.

REFERENCES

[1] Discovering interesting prediction rules with a genetic algorithm. In Proceedings of 1999 Congress on Evolutionary Computation (CEC' 99), pp. 13221329.
[2] Nguyen, D. and Widrow, B. (1990) Improving the Learning Speed of Two-Layer Networks by Choosing Initial Values of the Adaptive Weights.
[3] International Joint Conference on Neural Networks, San Diego, CA, III:21-26.
[4] Quinlan, J.R. (1986) Induction of Decision Trees. Machine Learning, 1, 81-106.
[5] Quinlan, J.R. (1987) Simplifying Decision Trees. International Journal of Man-Machine Studies, 27 , 221-234.
[6] Quinlan, J. R. (1996) Bagging, Boosting, and C4.5. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp. 725-730.
[7] V Adegoke, Daqing Chen, Ebad Banissi, and Safia Barikzai. Prediction of breast cancer survivability using ensemble algorithms. 2017.
[8] Abdulsalam Alarabeyyat, Mohannad Alhanahnah, et al. Breast cancer detection using k-nearest neighbor machine learning algorithm. In Developments in eSystems Engineering (DeSE), 2016 9th International Conference on, pages 35–39. IEEE, 2016.
[9] Dania Abed Aljawad, Ebtesam Alqahtani, ALKuhaili Ghaidaa, Nada Qamhan, Noof Alghamdi,Saleh Alrashed, Jamal Alhiyafi, and Sunday O Olatunji. Breast cancer surgery survivability prediction using bayesian network and support vector machines.In Informatics, Health & Technology (ICIHT), International Conference on, pages 1–6. IEEE, 2017.
[10] Hiba Asri, Hajar Mousannif, Hassa, AI Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83:1064– 1069, 2016.
[11] Bin Liu, Shanyi Wang, Qiwen Dong, Shumin Li, and Xuan Liu. Identification of dna-binding proteins by combining auto-cross covariance transformation and ensemble learning. IEEE transactions on nanobioscience, 15(4):328–334, 2016.
[12] Bin Liu, Shanyi Wang, Ren Long, and Kuo-Chen Chou. irspot-el: identify recombination spots with an ensemble learning approach. Bioinformatics, 33(1):35–41, 2016.
[13] Paritosh Pantola, Anju Bala, and Prashant Singh Rana. Consensus based ensemble model for spam detection. In Advances in Computing,Communications and Informatics (ICACCI), 2015 International Conference on, pages 1724–1727. IEEE, 2015.
[14] Ahmed Iqbal Pritom, Md Ahadur Rahman Munshi, Shahed Anzarus Sabab, and Shihabuzzaman Shihab"Predicting cancer recurrence using effective classification and feature selection technique" In Computer and Information Technology (ICCIT), 2016 19th International Conference on, pages 310–314. IEEE, 2016.