ISSN 2278-3091

Volume 9 No.2, March -April 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse04922020.pdf

https://doi.org/10.30534/ijatcse/2020/04922020

An Unsupervised approach for Predicting the Breast Cancer using K-Means with Compound Feature Generation

Rhagini A[1], Dhanush Narayanan T[2], Santhosh M[3], Sharmila C[4], Swetha G[5]

[1]Department of CSE, M.Kumarasamy College of Engineering, Karur, India, rhagini@gmail.com

[2] Department of CSE, M.Kumarasamy College of Engineering, Karur, India, dhanushruthresh2698@gmail.com

[3] Department of CSE, M.Kumarasamy College of Engineering, Karur, India, santhoshrainaa3@gmail.com

[4] Department of CSE, M.Kumarasamy College of Engineering, Karur, India, sharmi.0631@gmail.com

[5] Department of CSE, M.Kumarasamy College of Engineering, Karur, India, swethaganapathy13@gmail.com

ABSTRACT

Cancer is a serious threat and considered as most feared and dreaded disease. It is most common medical problem. In medical field, even experts are facing lots of problems and strain in both predicting and diagnosing the all types of cancer. There is numerous medical equipment available for diagnosing the cancer. Still medical experts are having puzzlement in predicting the cancer beforehand. So, we attempt to predict this deadly disease by using machine learning techniques. There are divers algorithms in machine learning for predicting the cancer. A few algorithms are k-nearest neighbor, support vector machine, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Bayesian decision rule, logistic Regression Model. We study the breast Cancer data obtained from the Breast Cancer Wisconsin (Prognostic) dataset (WPBC) from UCI machine knowledge with the hope of rising precise forecast replica for breast cancer using data mining system in this experiment. In WPBC dataset, the breast cancer dataset is categorized as twelve parameters. In order to improvise the efficiency, out of twelve parameters, the eight premium parameters should be chosen with the help of Compound Feature Generation (CFG). In CFG the feature selection and feature extraction are the two essential processes, which helps us to produce new feature subset from the available feature set. The newly obtained feature set has both elementary features and remodeled features for maintaining the orthogonality property. After applying CFG, the dataset is processed by k-means algorithm for better results. We evaluated various algorithms and found k-means algorithm has provided better accurate and efficient outcome.

Key words: Breast Cancer Prognosis, CFG, Feature Selection, K-means

1. INTRODUCTION

We have different forms and types of cancers such as lung cancer, blood cancer, skin cancer, stomach cancer, breast cancer etc., Biologically, cancer can be categorized into five: Carcinoma, Sarcoma, Melanoma, Lymphoma and Leukemia. There are many causes of cancer and depends on various factors including, biological, physical and lifestyle-related. The prognosis of cancer is much important because the probability of success is more when the cancer is predicted in the introductory stages. Census shows that among many types of cancer, the most threatening one is breast cancer which is the top causes of death among women. There is a lack of awareness about early prognosis and treatment which makes it more serious. We suggest a new technique by selecting the best features using Compound Feature Generation with the help of feature selection and feature extraction. We show experimentally that the genes chosen by our method give way improved organization presentation and are in nature pertinent to cancer. In difference with the baseline technique, our practices eradicate gene joblessness mechanically and give way improved and more dense gene subsets. Even using a smaller number of parameters, this algorithm gives us more precise results than the baseline method.

2. SUPPORT VECTOR MACHINES

Support Vector Machines is a classification algorithm in supervised machine learning techniques. It is a preferred technique because of its momentous accuracy and least computational power. P. S. Bradley and O. L. Mangasarian [1] has proposed DNA micro-arrays that authorize scientists to check thousands of genes concurrently and decide whether those genes are lively, agitated or quiet in regular or cancerous tissue. Because these new micro-array strategies make confusing quantity of raw data, new logical technique should be urbanized to sort out whether cancer tissues have characteristic signature of gene look over usual tissues or other kind of cancer tissues. In this paper, we lecture to the difficulty of assortment of a minute subset of genes from wide patterns of gene look data, evidence on DNA micro-arrays. Using obtained examples from cancer and usual patients, we construct a classifier suitable for hereditary diagnosis, as well as drug detection. R. Tibshirani and V. N. Vapnik al., [2] has proposed the holdup-vector network is a novel knowledge mechanism for two-group categorization evils. The mechanism theoretically equips the subsequent idea: input vectors are non-linearly charted to a very high dimension characteristic space. A special property of the choice surface ensures high sweeping statement aptitude of the knowledge mechanism. The idea at the back is support-vector system that has applied for the controlled case where the preparation data can be separated devoid of errors. We here make bigger product to non-separable preparation data. High sweeping statement aptitude of support-vector network use polynomial contribution alteration is established. We also contrast the presentation of the support-vector system to various traditional knowledge algorithms that has been developed for various purposes.

2.1 Regularization and Variable Selection

The goal of Support Vector Machine is to categorize the multi-dimensional data and to fit the training data set well. Also, it aims to avoid over fitting. Over fitting takes place when a function is too firmly fitted. The only solution to avoid over fitting is generalizing to new data points. Variable Selection is the method of choosing the creamy subset of predictors. Before variable selection, we need to find noises in data and if appropriate, we can add transformation of the variables. I. Guyon, J. Weston al., [3] explained the stretchy net with a new regularization and changeable assortment technique. Genuine planet data and a reproduction revise show that the stretchy net often outperforms the noose, while take pleasure in alike sparsely of depiction. The main objective of this paper is to rectify the problem of choosing the minimum subset from the available data. The author analyzed the training data of both cancer patients and normal patients and built a new classifier for genetic diagnosis. By using Recursive Feature Elimination in Support Vector Machine, the author illustrated an innovative method for selecting the gene. The author proved practically that the proposed method affords a more precise classification. In contrast to other methods, this paper showed that the method eradicates the redundancy of gene automatically and this leads to better results. However, the mechanism of regularization could not be understood apparently. And also, theoretical formulations could not be found.

3. NAÏVE BAYESIAN CLASSIFIER

Murat Karabatak [4] has proposed the simplest classifier algorithm called Naïve Bayesian for predicting the breast cancer. The principle of this algorithm is based on Bayes theorem. In this paper, the author proposed a new classifier in naïve bayes called Weighted Naïve Bayes (weighted NB). The author concluded the efficiency of Weighted NB is better than the regular NB. By using Weighted NB, features are fed to the inputs of Weighted NB classifier in the classification stage of breast cancer. In the training phase, the weights of the classifier are formerly tuned and this would be applied to the input features. The decision space can be obtained by running the weighted NB procedure. The author used some statistical parameters such as Sensitivity, Specificity and Accuracy to evaluate the Weighted NB classifier. Using different equations for each statistical parameter, the performance can be evaluated. The equation uses four common terms like True Positive, True Negative, False Positive and False Negative. Sensitivity can be calculated by taking ratio between TP and the sum of TP and FN. Specificity can be calculated by taking ratio between TN and the sum of TN and FP. Accuracy can be calculated by taking ratio between sum of TP and TN and the sum of all the four. The author mentioned that a grid search mechanism was used to find optimal results. But the inconvenience in using this algorithm is, that the grid search mechanism is computationally overpriced.

4. RANDOM FOREST CLASSIFIER

The Random Forest (RF) algorithm is the combination of several decision trees and it is widely used in classification methods. Trees are generated using binary tree concept and it is recursively partitioned for further operations. Cuong Nguyen, Yong Wang, Ha Nam Nguyen [5] proposed the random forest algorithm in breast cancer prediction. In this paper, the authors came up with two-phase method. In the phase one, the training set and the validation set are trained and tested to pick the best features. Feature Ranking is the essential procedure in phase one. By using feature ranking, all the features have been sequenced in ascending order. To improve the classification accuracy, the selected features from phase one has been valued in phase two. The authors proposed this paper in a four-step classification: The first step is to train learning algorithm using n-fold cross validation. The second step is estimating the Bayesian probability and feature ranking value. In the third step, the redundant features should be deleted by Backward Elimination for improving classification accuracy. The fourth step is to iterate the above steps until the stopping criteria met. In this paper, number of parameters has been determined such as number of trees and number of features. This paper showed 25 trees in random forest and 15 features. The difficulty in using this algorithm is that the proposed method has a greater number of trees.

5. K- NEAREST NEIGHBOR ALGORITHM

K-NN is a classification algorithm which calculates various distances from each element to all other elements. This algorithm is mainly based on prime parameter k that represents the number of neighbors selected for assigning the class to the new element and the choice of the distance. Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou [6] has proposed a paper on predicting the breast cancer using K-NN algorithm. In this paper, the authors analyzed numerous distances and found different values of the unique parameter k, using classification rules in K-NN. The efficiency of this algorithm chiefly depends on the distance and the value of parameter k. A distance in the space should satisfy four axioms: Non-negativity, Symmetry, Reflexivity, Triangle Inequality. Different distances have been discussed and calculated in this paper. The distances are Cityblock distance: 1-distance, Euclidean distance: 2-distance, Minkowski distance: p-distance, Tchebychev distance: ∞-distance, Cosine distance and Correlation distance. All these distances are used to find the class of a new element. In this paper, the k-value is chosen to minimize the classification error and k-value should be greater. Like cross-validation, we have various heuristic techniques to choose the better value for k. Various classification rules has been used such as nearest rule, random rule and consensus rule. Though these distances gave best results, it is still time consuming.

6. K-MEANS ALGORITHM

The above discussed algorithms come under the supervised machine learning technique. In supervised learning, both the input and output data will be given in the training and test dataset. There may be no chance for learning or discovering hidden features since it is constrained to the class label. So, another type of machine learning is unsupervised. Here only the input data will be fed. Learning from the input dataset, along with the hidden features, the output will be found which leads to better efficiency. One of the unsupervised learning algorithms is K-means clustering technique. The main purpose of clustering is to identify the valuable insights since it can generalize to different shapes and sizes of clusters. The initial step of k-means is to cluster the datasets based on similarities and differences. The concealed patterns can be perceived with the help of those clusters. Jyotsna Nakte and Varun Himmatramka [7] has proposed the paper on predicting the breast cancer with K-means algorithm. In this paper, the authors stated that this algorithm can be favorably utilized by considering huge portions of valuable issues. Initially, the objects in the datasets were clustered randomly or with some preceding knowledge. The distance between the different clusters and between the data objects in the clusters were calculated and analyzed. With the calculated distance, the mean of every cluster should be found. The steps will be repeated until the mean of each clusters cannot be improvised anymore. Prediction will be better when hidden features are addressed. Thus, it can be achieved by using K-means algorithm.

7. FEEDFORWARD NETS FOR INTERPOLATION AND CLASSIFICATION

Feedforward nets have the capability of computing certain interconnections of simple neurons. These are formulated in a layered network. Each neuron calculates a scalar function of its aggregate input. These interconnections of processors are known as feedforward neural nets. E. D. Sontag al., [8] has proposed a paper about feedforward neural nets. To preserve parallel relationship of the input given to any neuron, aggregate the outcomes from the processors which are connected. They are weighted depending upon their real-valued coefficients. The outcome of the nets can be obtained from the output of the final neuron. The function calculated by the nets is based on three important parameters: weights, interconnection pattern and the selection of the scalar function, denoted as θ. The function by nets may seems like burdensome process. To reduce efforts, one can bypass the intermediate layers for the blunt connections from the inputs and results. In this paper, the author completely concerned with three-layer nets. However, the neurons that are present between the first and final nodes are hidden, known as hidden-units. The author stated about Heaviside function, which can have value as 1 for positive parameters and 0 for others.

8. BREAST CANCER MINING

The purpose of mining is to extract knowledge from the raw data. Also mining helps us to categorize the data into erroneous data and valuable data. Finding relationships between data is even better when mining occurs. R.Preetha and S.Vinila Jinny.,[9] has proposed a paper for breast cancer prediction using data mining. The authors stated that usage of data mining algorithm in breast cancer prognosis leads to less time consumption. This is because, the data mining algorithms can process huge medical dataset in few seconds. Considering different parameters in multidimensional view, the data mining method will give results. Clipping the worthwhile data from the provided dataset involves discovery. It is carried out in following iterative steps: Data cleansing, Data integration, Data desire, Data transformation, Facts mining, Pattern assessment and Expertise example. The main purpose of mining in breast cancer prognosis is that, processing the impurities, which is a serious threat to accuracy in prognostic models. Those impurities are: Missing data, Outliers and Imbalanced data. The authors also stated many attributes for breast cancer such as Morphology, Topography, Survival status etc., The author stated that mining is the mandatory step to be taken before using any algorithm for cancer prognosis.

9. ENDOMETRIAL CANCER RELATED TO BREAST CANCER

Cancer cells have the ability to multiply in numbers rapidly. It can be spread to whole body cells and can create cancer in various parts also. Endometrial cancer is a uterine cancer, that has the capability to invade the cells in other parts of the body. It can occur by means of lymph or via blood. C. Cortes and V. N. Vapnik al., [10] has proposed that this study dealt with the computational capabilities of certain interconnections of simple neurons called processors. These are concurred in a covered network, such that each computer is scheming a scalar purpose that is the start or response function of its collective contribution. These interconnections are commonly called as feed forward neural net. They have paying attention as a potentially helpful replica of similar calculation. The contribution fed to any known computer is a bi mixture of the production of all the processors that attach to it, biased according to real-valued coefficients. The production of the last computer is in use as the production of the net. The weights are connected with the interconnection pattern and are used for determining completely the function computed by the net.

Base on global Federation of Gynecologists and Obstetricians (FIGO) criteria, endometrial cancer is surgically theatrical. Despite this strategy, presentation of total surgical staging for endometrial cancer is contentious. The GOG [Gynecologic Oncology Group] surgical physical explains the complete surgical performance of endometrial cancer as taking away of the breast cancer, cervix, adnexa, and pelvic and armpit lymph node tissues, and get pelvic washings. GOG describes pelvic lymphadenectomy as taking away of the nodal tissue from the distal semi of the ordinary iliac arteries, the forward and medial feature of the proximal half of the outside iliac blood vessel and vein, and the distal semi of the obdurate fat protection frontal to the obdurate nerve; armpit lymph lump analysis is described as elimination of nodal tissue over the distal lesser vena cava from the height of the lesser mesenteric blood vessel to the middle right ordinary iliac road and elimination of the nodal tissue flanked by the aorta and absent ureter from the mid lesser mesenteric route to the mid left ordinary iliac road.

Thus, a number of practitioners may opt for discriminating lymph node. Though, display data propose that patients who experience manifold site example had better continued survival over those who had incomplete or no example carry out. The warning to nodal example versus filled analysis is that examination or palpation of nodes has not been revealed to be a responsive process for discovery of optimistic lymph nodes, with less than 10% of patients with lymphadenopathy encompass disgustingly concerned nodes.

10. CONCLUSION

This paper has discussed about testing different algorithms. The final outcome of the research targeted on correctness of the algorithms in the training dataset. It depended on Breast Cancer Wisconsin (Prognostic) data set (WPBC). This work chose K-SVM, Random Tree, K-Nearest Neighbor, Feedforward neural nets, Naive Bayesian and K-means algorithm for test; the result specified which one is the best to detect breast cancer. The test result shows that the K-means can be a better choice of algorithm in the case of detailed prediction. Though the accuracy rate is predominantly higher in many classification algorithms like SVM, K-NN. The efficiency rate can be achieved by prognosing the concealed features with the help of unsupervised learning technique. So, K-means can be used for predicting the breast cancer effectively with the help of Compound Feature Generation to improve its accuracy rate.

REFERENCES

1. P. S. Bradley and O. L. Mangasarian, **“feature selection via concave minimization and support vector machines,”** in machine learning: proceedings of the fifteenth international conference (icml ’98). Morgan kaufmann, san francisco, 1998, pp. 82–90.

2. R. Tibshirani, **“Regression shrinkage and selection via the lasso,”** Journal of the Royal Statistical Society, vol. 58(1), 1996, pp. 267-288
https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

3. I. Guyon, J. Weston, and S. Barnhill, **“Gene selection for cancer classification using support vector machines,”** Machine Learning, vol. 46, pp. 389–422, 2002.
https://doi.org/10.1023/A:1012487302797

4. Murat Karabatak, **“A new classifier for Breast Cancer Detection based on naïve bayesian”,** Elsevier**,** August 2015,pp.32-36

5. Cuong Nguyen, Yong Wang, Ha Nam Nguyen**, “Random Forest Classifier combined with feature selection for breast cancer diagnosis and prognostic ”,** Journal of Bio-medical Science and Engineering – January 2013, pp.551-560

6. Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou **“Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules”,** International Journal of Computer Application, January 2013, pp.1-5
https://doi.org/10.5120/10041-4635

7. Jyotsna Nakte and Varun Himmatramka, **“Breast cancer prediction using data mining techniques”,** International Journal on Recent and Innovation Trends in Computing and Communication, Vol:4, Issue: 11, November 2016,pp.55-

8. E. D. Sontag al, **“Feedforward Nets for Interpolation and classification”**, Journal of Computer and System Sciences 45 , 1992, pp. 20-48

9. R.Preetha and S.Vinila Jinny, **“A Research on Breast cancer prediction using data mining techniques”**, International Journal of Innovative Technology and Exploring Engineering, Vol:8, Issue:11S2, September 2019, pp.362-370

10. C. Cortes and V. N. Vapnik**, “Support vector networks,” Machine** Learning, vol. 20, 1997.pp.273-297
https://doi.org/10.1007/BF00994018

11. S.Thilagamani, N.Shanthi, **“Literature survey on enhancing cluster quality”**, International Journal on Computer Science and Engineering Vol. 02,No. 06, 2010,pp1999-2002

12. E.T.Venkatesh, P.Thangaraj, S.Chitra, **"Classification of cancer gene expressions from micro-array analysis"**, International Conference Innovative Computing Technologies (ICICT), 2010,pp.1-5

13. P. Pandiaraja, N Deepa 2019 , **A Novel Data Privacy-Preserving Protocol for Multi-data Users by using genetic algorithm** , Journal of Soft Computing , Springer , Volume 23 ,Issue 18, Pages 8539-8553.

14.K Sumathi, P Pandiaraja , **Dynamic alternate buffer switching and congestion control in wireless multimedia sensor networks** , Journal of Peer-to-Peer Networking and Applications , Springer

15. Dr.A.Nagarajan, J. Vasanth Wason, **"Machine Learning Approach to Predict Lung Cancer using CT scan Images",** International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Volume 8, No.6, November – December 2019, pp.2974-2976. https://doi.org/10.30534/ijatcse/2019/48862019

16. Dr. Brijesh Kumar Bhardwaj, **"A Critically Review of Data Mining Segment: A New Perspective",** International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Volume 8, No.6, November – December 2019,pp.2984-2987. https://doi.org/10.30534/ijatcse/2019/50862019

17.P.RajeshKanna and P.Pandiaraja 2019, **An Efficient Sentiment Analysis Approach for Product Review using Turney Algorithm** , Journal of Procedia Computer Science , Elsevier ,Vol 165 ,Issue 2019, Pages 356-362 https://doi.org/10.1016/j.procs.2020.01.038