International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Implementation of Tumor Prediction System using Classification Algorithms

B. Kranthi Kiran

Abstract: ***As the huge volume of healthcare data was being unused, recent researchers were focused on predicting the many diseases by analyzing the past patient records. In continuation with that, there are lot of researches focused on predicting the tumor on the human body. In this research, two widely used classification algorithms called Naïve Bayes and Random tree were considered for implementation and analysis with the UCI Machine learning Tumor data set. The data cleaning technique called "Replace Missing Values" in the WEKA tool has been considered for cleaning the data. The implementation has been done with the original dataset and the cleaned dataset. Finally, it is found that the Random tree algorithm is performed well with improved accuracy and reduced error rate. The accuracy obtained before data cleaning is 90.8333% and after data cleaning is 93.3333 %. Similarly, the error rates were reduced reasonably and they are 9.1667 % before data cleaning and 93.3333 % after data cleaning. In future, the data cleaning techniques has to be tuned well to improve the accuracy further.***

KEYWORDS: Data Cleaning, Naïve Bayes Algorithm, Random Tree Algorithm, Tumor Prediction and Classification.

I. INTRODUCTION

The large volume of healthcare data was unused and the revolution of machine learning enables the healthcare data to be analysed for prediction. Many researchers focus on analysing the various diseases of human beings. This research paper is focusing on the prediction of the tumor in the human body by considering the history of patient records. The data mining tool called WEKA has been used for implementation and analysis. This tool has the inbuilt feature of many classification algorithms and data pre-processing techniques. This paper not only helps the patients in identifying the tumor in the earlier stages, also helps the medical professionals to identify and conclude the occurrence of tumor, its location, size and type etc., This implementation also reduces the unnecessary expenses of the patients by eliminating the requirements of undergoing various medical tests. It is a data mining task of predicting the value of categorical variable (target or class). This is done by building a model based on one or more numerical or categorical variables. Classifier, Classification model, Feature, Binary Classification, Multi-class classification, Multi-label classification are the basic terminologies used in classification algorithms. Electronic mail spam classification, Bank clients lend pay readiness prediction, cancer cell credentials, Sentiment analysis, Drugs classification, Facial key points detection, Pedestrian detection in automotive car driving are the applications of classifications. The risks ofthe tumor in human body is discussed here and they are a swelling of a part of the body generally without inflammation caused by an abnormal growth of tissue whether benign or malignant. The Innovative age is the most significant risk factor. Many risk factors are related to a person's life style or overall health. Excess body weight contributes to 1 in 5 cancer related deaths in the U.S. Obesity causes breast, colorectal and other types of cancers. Carcinogens a substance that cause cancer by damaging DNA. Tobacco especially second-hand tobacco is a proven cause of cancer. The more a person drink's alcohol, the more likely they are to develop cancer. The risk is higher for those who drinks alcohol and also use tobacco. Viruses like hepatitis A and hepatitis B causes cancer. Lack of exercise, poor health condition, habits and hormones are also causing cancer in human body.

II. LITERATURE REVIEW

In humanoid body the most significant organ is the intellect. Savor, emotions, earshot, tad, view etc., everything are controlled by brain. In every facets of medical field brain plays a significant role. One of the dangerous diseases in the last decade is brain tumor. Brain tumor prediction is a very difficult process. using some segmentation and classification techniques of data mining finding the brain tumors in patient's body is explained by proposed system. An author used K-means clustering algorithm for segmentation. Further an author used SVM classification and decision tree algorithms to perform classification of MRI brain image and to predict better classification algorithm technique and using this extract the parts of the tumor in brain[1].

In females one of the deadliest tumors is breast tumor. In breast tissue when the development of the cells become out of control it happens. Malignant and benign are the two types of tumors. To classify the type of the tumor classification algorithms are used. Using voting classifier method, the comparison of the results of supervised classification algorithms and grouping of these algorithms is the key of this article. For enhanced classification one of the ensemble methodologies where we can combine various models is voting. Combination of algorithms should be taken astutely in future without over fitting problem[2].

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
B. Kranthi kiran*, Associate Professor, JNTUHCEH. Kukatpally, Hyderabad, Mail: kiranjntuh321@gmail.com.

Using a temperature silhouette to prognosis anonymous thermal and geometrical parameters of a tumor area on the skin surface of diverse organs which may be obtained by ultraviolet thermography an inverse process is offered in this paper. To guess the key parameters, such as deepness, temperature generation rate, and tumor size the GA was used. The reliability between the actual and foretold parameters is shown by the obtained outcomes. For the estimate problem it was indomitable that a GA-based approach is well appropriate[3].

Based on brain MRI pictures the era of hominids is predicted by this article. Two dissimilar approaches are suggested to excerpt T1-MRI features then ELM is hired to guess era. The time needed to era valuation is low, the suggested technique has satisfactory presentation[4].

With a time of melodramatically augmented use of electronic medicinal registers and diagnostic imaging. The terrific achievement of MLA at image recognition tasks in modern ages intersects. The MLA as applied to medicinal image investigation, concentrating on CNN, and highlighting clinical aspects is hosted by this review[5].

A significant problem in the tumor detection procedure is forecasting tumor sickness state. Variation SVM of MLAs usage is projected by an author and assess them on actual clinical information associated to thyroid tumor, colon tumor and liver tumor. The virtuous presentation of support vector machines is displayed by untried outcomes[6].

With data mining methods a survey on medicinal image feature miscellany is offered in this article.To predict the tumor disease at former stage, Thus the survey helps to recognize the data mining methods. For the study ANN, decision tree, Support Vector Machine, Back propagation, Naïve bays and Regression are deliberated. Based on presentation, rapidity, cost and accurateness numerous outcomes have given by different algorithms. A better technique can be found out in future to predict the cancer disease with progresses in current techniques[7].

One of the testing and vital steps to cure preparation is uncovering and dissection of irregularities existing in the liver which enlarges the existence of a patient. As side effects cannot be eminent even the cancer is in its progressive phase liver cancer increases the demise rate. For noticing and categorizing the US imageries a new scheme has been discoursed in this article. For the liver cancer recognition and cataloguing system, the suggested method is well suitable[8].

Wrecking the opposed surrounding cells and their regular tasks and some cells in the body emerging aberrantly are the condition to cause cancer. Cancer feasts generally in all parts of the humanoid body. To forecast the protein causing tumor varied info excavating schemes are used in this article. For the diagnosis of tumor and to discover the medicine for the protein causing tumor the protein forecast would help the medicinal region[9].

Surmising the knowledge from bulky quantity of datasets is KDD. On the dataset of breast tumor an author used J48 KDD algorithm in this article. For C4.5 algorithm J48 is a java execution provided by Weka tool. Based on the attributes such as node-caps, era, lump size, menopause, illuminate etc to predict the repeated proceedings an author scrutinizes the verdict tree created by the algorithm using 10-fold irritable authentication technique[10].

Tumor at 12.4% of all demises is the second major reason of demise in Palestine. One of the stimulating determinations of emerging a medicinal information excavating applications is used for forecasting the survivability of a sickness. To forecast the survivability of tumor patients 2 cataloguing replicas are practical on tumor patient's records. By decision tree and naive Bayes algorithms an author build further classifiers in upcoming[11].

Radiologists demeanor FNA technique of breast cancer for breast tumor diagnosis in patients. Topographies such as lump radius, feel and fractal sizes are publicized by this technique. By dataset from Wisconsin Breast Tumor Information using sophisticated classifiers forecast breast tumor is the goal of this article[12].

By KDD methods that are recycled in today's medicinal investigation information excavating proposes to endow through a methodical survey of present methods of knowledge detection in databases. Depends on current software metrics with KDD methods to assistance developers recognize flaws and here by progress the software excellence is the key impartial of this article. For lung tumor forecast cataloguing and bunching and connotations excavating of data mining methods are discussed by an author in this paper. How data mining techniques are used for software flaw forecast are also discoursed by an author[13]. Heart disease prediction systems were well analyzed using classification algorithms.[14-16].

III. PROPOSED SYSTEM AND DATASET

The data was collected from the University of Medical centre, Yugoslavaia. The dataset name was Primary Tumor. There are totally 339 instances and out of these 20 instances were selected for the implementation. These instances have 18 attributes including the class label. All attribute values in the database have been entered as numeric values corresponding to their index in the list of attribute values for that attribute domain as given below. The attributes are class label, age, sex, histologic-type, degree-of-diffe, bone, bone-marrow, lung, pleura, peritoneum, liver, brain, skin, neck, supraclavicular, axillar, mediastinum and abdominal. As the datasets were subjected to missing values, the data cleaning technique called "Replace Missing Values" has been implemented through WEKA tool. The Naïve Bayes algorithm and Random Tree has been implemented with the collected tumor dataset in two stages. The two stages are (I) Before data Cleaning and (ii) After Data Cleaning. The data cleaning technique called "ReplaceMissingValues" in WEKA tool has been applied. The results obtained in both the stages were compared for analysis. The parameters considered for comparison are the number of records classified correctly, number of records classified by the classifier incorrectly, Kappa statistic, Meana absolute error, Root mean squared error, root absolute error and Root relative squared error. The proposed System of Tumor Prediction is shown in Fig.1. The tumor dataset is fed as input to Naïve Bayes and Random tree algorithms before and after data cleaning. The results obtained were recorded for analysis.

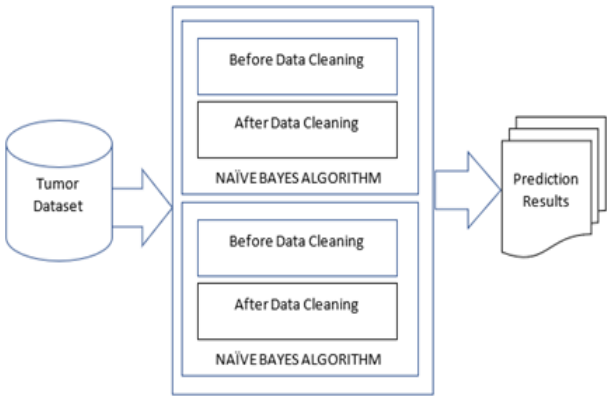

Fig.1. Proposed System for Tumor Prediction

IV. IMPLEMENTATION AND RESULTS

A. Naïve Bayes Algorithm With Tumor Dataset Before Data Cleaning

The Naïve Bayes algorithm implemented with Tumordatasetbefore data cleaning yields the following results. The accuracy obtained is 69.1667 % with the total number of correctly classified instances as 83. The error rate obtained is 30.8333 % with the incorrectly classified instances of 37. The total number of records considered is 120. The other parameters and the values are kappa statistic of 0.6632, Mean absolute error of 0.0337, Root mean squared error of 0.181, Relative absolute error of 33.765 % and Root relative squared error of 81.2794 %. The results yield has been given in the Table 1.

Table 1. Results Obtained with Naïve Bayes with Tumor Dataset Before Data Cleaning.

Parameter	Values Obtained	Accuracy and Error rate
Correctly Classified Instances	83	69.1667 %
Incorrectly Classified Instances	37	30.8333 %
Kappa statistic	0.6632	
Mean absolute error	0.0337	
Root mean squared error	0.181	
Relative absolute error	33.765 %	
Root relative squared error	81.2794 %	
Total Number of Instances	120	

B. Naïve Bayes Algorithm With Tumor Dataset After Data Cleaning

Table 2. Results Obtained with Naïve Bayes with Tumor Dataset after Data Cleaning.

Parameter	Values Obtained	Accuracy and Error rate
Correctly Classified Instances	87	72.5%
Incorrectly Classified Instances	33	27.5%
Kappa statistic	0.6967	
Mean absolute error	0.0338	
Root mean squared error	0.1546	
Relative absolute error	33.8262%	
Root relative squared error	69.4374 %	
Total Number of Instances	120	

The Naïve Bayes algorithm implemented with Tumor dataset after data cleaning yields the following results. The accuracy obtained is 72.5 % with the total number of correctly classified instances as 87. The error rate obtained is 27.5 % with the incorrectly classified instances of 33. The total number of records considered is 120. The other parameters and the values are kappa statistic of 0.6967, Mean absolute error of 0.0338, Root mean squared error of 0.1546, Relative absolute error of 33.8262 % and Root relative squared error of 69.4374 %. The results yield has been given in the table 2.

C. Random Tree with Tumor Dataset Before Data Cleaning.

The Random Tree algorithm implemented with Tumor dataset before data cleaning yields the following results. The accuracy obtained is 90.8333 % with the total number of correctly classified instances as 109. The error rate obtained is 9.1667 % with the incorrectly classified instances of 11. The total number of records considered is 120. The other parameters and the values are kappa statistic of 0.8977, Mean absolute error of 0.0141, Root mean squared error of 0.0842, Relative absolute error of 14.1479 % and Root relative squared error of 37.8036 %. The results yield has been given in the table 3.

Table 3. Results Obtained with Random Tree by Tumor Dataset before Data Cleaning

Parameter	Values Obtained	Accuracy and Error rate
Correctly Classified Instances	109	90.8333%
Incorrectly Classified Instances	11	9.1667%
Kappa statistic	0.8977	
Mean absolute error	0.0141	
Root mean squared error	0.0842	
Relative absolute error	14.1479 %	
Root relative squared error	37.8036 %	
Total Number of Instances	120	

D. Random Tree With Tumor Dataset After Data Cleaning

The Random Tree algorithm implemented with Tumor dataset after data cleaning yields the following results. The accuracy obtained is 93.3333 % with the total number of correctly classified instances as 112. The error rate obtained is 6.6667 % with the incorrectly classified instances of 8. The total number of records considered is 120. The other parameters and the values are kappa statistic of 0.9254, Mean absolute error of 0.0347, Root mean squared error of 0.0956, Relative absolute error of 34.7954 % and Root relative squared error of 42.9327 %. The results yield has been given in the table 4.

Table 4. Results Obtained with Random Tree by Tumor Dataset after Data Cleaning.

Parameter	Values Obtained	Accuracy and Error rate
Correctly Classified Instances	112	93.3333 %
Incorrectly Classified Instances	8	6.6667 %
Kappa statistic	0.9254	
Mean absolute error	0.0347	
Root mean squared error	0.0956	
Relative absolute error	34.7954 %	
Root relative squared error	42.9327 %	
Total Number of Instances	120	

The comparison of Accuracies and Error rate in both the cases of before and after data cleaning with the Naïve Bayes Classification algorithm is shown in table 5 and figure 2. It is found that the results were better after the data cleaning has been applied for the dataset.

Table 5. Results obtained with Naïve Bayes Algorithm

Parameter	Before Data Cleaning	After Data Cleaning
Accuracy	69.1667 %	72.5%
Error Rate	30.8333 %	27.5%

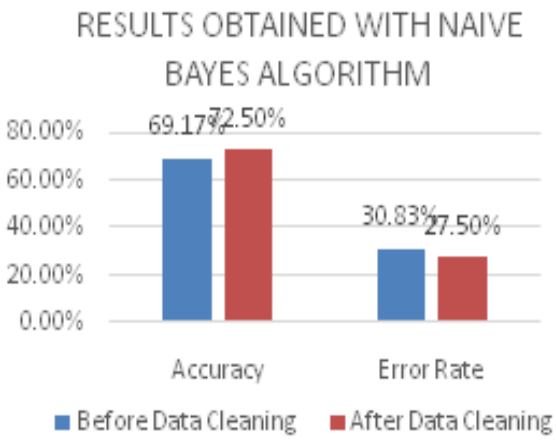

Fig.2. Results obtained with Naïve Bayes Algorithm

The comparison of Accuracies and Error rate in both the cases of before and after data cleaning with the Random Tree Classification algorithm is shown in table 6 and figure 3. It is found that the results were better after the data cleaning has been applied for the dataset.

Table 6. Results obtained with Random Tree Algorithm

Parameter	Before Data Cleaning	After Data Cleaning
Accuracy	90.8333%	93.3333 %
Error Rate	9.1667%	6.6667 %

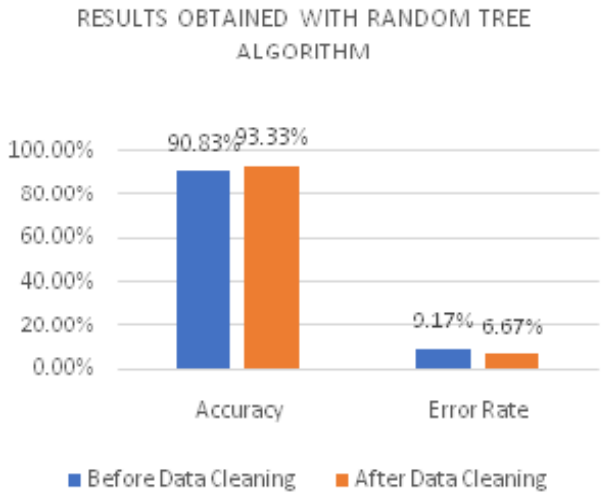

Fig.3. Results obtained with Random Tree Algorithm

V. CONCLUSION

After a deep study on the classification algorithms for implementing in tumor prediction, the two classification algorithms called Naïve Bayes and Random Tree has been chosen for implementation. The datasets were collected from the Machine learning database which has been widely used. The implementation has been carried out in two stages and they are before and after data cleaning. The results obtained were concluded that**,** the Random tree algorithm is performed well with improved accuracy and reduced error rate. The accuracy obtained before data cleaning is 90.8333% and after data cleaning is 93.3333 %. Similarly, the error rates were reduced reasonably and they are 9.1667 % before data cleaning and 93.3333 % after data cleaning. In future, the data cleaning techniques has to be tuned well to improve the accuracy further. The Naïve Bayes algorithm yields 72.5 % of accuracy and 27.5% error rate after data cleaning. The same algorithm yields 69.1667 % of accuracy and 30.8333% error rate before data cleaning technique.

REFERENCES

1. Deepak, K. S., Gokul, K., Hinduja, R., & Rajkumar, S. (2013, February). An efficient approach to predict tumor in 2D brain image using classification techniques. In 2013 International Conference on Information Communication and Embedded Systems (ICICES) (pp. 559-564). IEEE.
2. Kumar, U. K., Nikhil, M. S., & Sumangali, K. (2017, August). Prediction of breast cancer using voting classifier technique. In 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM) (pp. 108-114). IEEE.
3. Hossain, S., Abdelaal, M., & Mohammadi, F. A. (2016). Thermogram assessment for tumor parameter estimation considering body geometry. Canadian Journal of Electrical and Computer Engineering, 39(3), 219-234.
4. Afshar, L. K., & Sajedi, H. (2019, January). Age Prediction based on Brain MRI Images using Extreme Learning Machine. In 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) (pp. 1-5). IEEE.
5. Ker, J., Wang, L., Rao, J., & Lim, T. (2017). Deep learning applications in medical image analysis. Ieee Access, 6, 9375-9389.
6. Turki, T. (2018, March). An empirical study of machine learning algorithms for cancer identification. In 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC) (pp. 1-5). IEEE.
7. Saranya, P., & Satheeskumar, B. (2016). A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques. International Journal of Computer Science and Mobile Computing, 5(5), 713-719.
8. Rajesh, G. (2018). Liver cancer detection and classification based on optimum hierarchical feature fusion with PeSOA and PNN classifier, Bio Medical Research, 29(1).
9. Lobo, S., & Pallavi, M. S. (2018, April). Predicting Protein in Cancer Diagnosis Using Effective Classification and Feature Selection Technique. In 2018 International Conference on Communication and Signal Processing (ICCSP) (pp. 156-159). IEEE.
10. Bhargava, N., Sharma, S., Purohit, R., & Rathore, P. S. (2017, October). Prediction of recurrence cancer using J48 algorithm. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES) (pp. 386-390). IEEE.
11. Alhaj, M. A., & Maghari, A. Y. (2017, May). Cancer survivability prediction using random forest and rule induction algorithms. In 2017 8th International Conference on Information Technology (ICIT) (pp. 388-391). IEEE.
12. Sharma, A., Kulshrestha, S., & Daniel, S. (2017, December). Machine learning approaches for breast cancer diagnosis and prognosis. In 2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp) (pp. 1-5). IEEE.
13. Periasamy, A. P., & Arutchelvan, K. (2017). Data Mining Techniques in Multiple Cancer Prediction. International Journal of Advanced Research in Computer Science and Software Engineering, 7(5), 472-475.
14. Jothikumar, R. (2016). C4. 5 classification algorithm with back-track pruning for accurate prediction of heart disease.
15. Jothikumar, R., Susi, S., Sivakumar, N., & Ramesh, P. S. (2018). Predicting life time of heart attack patient using improved C4. 5 classification algorithm. *Research Journal of Pharmacy and Technology*, *11*(5), 1951-1956.
16. Jothikumar, R., Sivabalan, R. V., & Sivarajan, E. (2015). Accuracies of j48 weka classifier with different supervised weka filters for predicting heart diseases. *ARPN J Eng Applied Sci*, *10*, 7788-7793.

AUTHORS PROFILE

Dr. B. Kranthi Kiran, is working as the Associate Professor in the department of Computer Science & Engineering, JNTUH College of Engineering, Hyderabad. He has published many Scopus indexed Journals, participated in various workshops and conferences.