ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 8, Issue, 3, pp. 15777-15789, March, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

CANCER DETECTION & PREDICTION USING DUAL HYBRID ALGORITHM

Kritharth Pendyala., Divyesh Darjee., Vaishnavi Adhyapak and Vaishali Gaikwad

Department of Information Technology K.J.S.I.E.I.T. Mumbai, Maharashtra, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0803.0004

ARTICLE INFO

Article History:

Received 16th December, 2016
Received in revised form 25th January, 2017
Accepted 23rd February, 2017
Published online 28th March, 2017

Key Words:

Breast Cancer detection, prediction, Hybrid, Naïve Bayes, Support Vector Machine, classifiers, Machine Learning, Diagnosis, Accuracy, ductal carcinoma

ABSTRACT

Breast Cancer has been a leading cause of death among women across the globe; meanwhile, it is a type of cancer that comes under the treatable category. Although, early diagnosis and accurate detection of this disease promises an extended life-cycle and long survival of the diagnosed patients. The algorithms that are used in this work are Naïve Bayes and Support Vector Machine (SVM). SVM is a part of machine learning classification which uses labels to cluster data. These classifier algorithms are compared based on the performance factors i.e. accuracy of classification and execution time required. From the experimental hybrid algorithm we will overcome the disadvantages that are currently present when computing the prediction with Naïve Bayes classifier. Naïve Bayes has certain drawbacks that affects clustering and classification of data. Data scarcity, unavailability of contiguous data sources and strong presumptions of the shape of data distribution are the primary factors that undermine the utilities of Naïve Bayes classifier. The basic idea while deploying a dual hybrid algorithm is to ensure that there is a much better efficiency and that a single instance can effectively predict the disease unlike in a single algorithm system.

INTRODUCTION

Ductal Carcinoma is a potentially fatal disease caused primarily by environmental factors that constantly mutate genes that encode critical cell-regulatory proteins. The resultant aberrant cell behaviour leads to expansive masses of abnormal cells that destroy surrounding normal tissue and can spread to vital organs resulting in disseminated disease, commonly a harbinger of imminent patient death. More significantly, globalization of unhealthy lifestyles, particularly cigarette smoking and the adoption of many features of the modern western diet (high fat, low fibre content) will increase occurences of this disease.

Data mining techniques involve the use of sophisticated data analysis tools and classification methods to discover previously undisclosed, valid patterns and relationships in a large data set. These tools can include statistical models, mathematical algorithms and machine learning methods in early detection of ductal carcinoma. In classification mining, the learning scheme is fed with a set of classified examples from which the system is taught to train itself to classify unknown examples of clusters. In association learning, any association among features is sought, not just the ones that predict a particular class value. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be generated is not a discrete class but a whole quantity. In this study, to classify the data and to mine certain patterns from the input dataset is the objective of the algorithm. Naïve bayes and Support Vector Machine are used in a dual effective manner to give the appropriate results. Data Mining techniques are implemented together to create a novel method to diagnose the existence of ductal carcinoma for a particular patient.

When beginning to work on a data mining problem, it is extremely important to bring all the data together into a set of instance classes. Integrating and unifying data from different sources usually puts-forth several challenges. The data must be assembled, integrated, and cleaned up. Then only will it be used for processing through machine learning techniques and association mining. This developed system can be used by physicians and patients alike to easily know the status of the health of a patient suffering from ductal carcinoma, without physically screening them for the disease. Also it is useful to record and save large volumes of sensitive information which can be used to gain knowledge about the disease and its treatment.

Section II denotes and indicates the work that has been done in the current sector and field of cancer detection. Algorithms chosen have primarily been to overcome shortcomings of the previous work done.

Corresponding author: **Kritharth Pendyala*
Department of Information Technology K.J.S.I.E.I.T. Mumbai, Maharashtra, India

Section III depicts the various algorithms deployed to ensure the proper running if the current system in an efficient and effective manner.
Section IV includes the description of the list of modules required to develop and deploy the proposed system.

Section V represents the applications of the proposed system, this includes the fields pertaining mainly to the healthcare industry

Related Work

An analysis of Classification and Prediction data mining techniques is provided in "Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset"[1]. These results are based on the selected Wisconsin University Breast Cancer dataset. The data was chosen in terms of randomness.

The Naïve Bayes classification algorithm shows that the success rate is around 85-90 percent and the error rate is around 10-15 percent. The prediction algorithm also shows the same success and error rates[2]. The algorithms can be improved for various small instances that will help in increase of the success rate. Analysis of the efficiency on their work shows that there is a scope for improving the success rate and to help recess the error rate by employing the techniques of machine learning like Support Vector Machine algorithm.

The main objective of the existing system is to provide an analytical study of the efficiency and accuracy of Naïve Bayes classifier in prediction of cancer. Using Support vector machine algorithm[3] will give us an opportunity to overcome the shortcomings faced by prediction using Naïve Bayes classifier, such that we have the best cancer detection and prediction system.

Krishnaiah and Aruna Sundaram developed a hybrid statistical pattern recognition algorithm with a title of," Hybrid SPR algorithm to select predictive genes for effectual cancer classification" [4]. In this study, a hybrid Statistical pattern recognition algorithm has been proposed to reduce the dimensionality and select the predictive genes for classification of cancer. Colon cancer gene expression profiles having 62 samples of 2000 genes were used for the experiment. The disadvantage here is that a minimum of 62 samples were required to accurately predict cancer.

Bijan Moghimi-Dehkordi and Azadeh Safaee conducted experiments on survival rates and prognosis in Asia with a title of "An overview of colorectal cancer survival rates and prognosis in Asia" [5]. In their research the single algorithmic outputs could only show certain outputs which were accurate upto the 90%, the rest were found to have lesser accuracy and more error rates. In this study, in Asia, the overall cure rate of colorectal cancer has not improved dramatically in the last decade, 5-year survival remaining at approximately 60%.

Srinivas Mukkamata *et al*. found that Computational intelligent technique [6] that can be useful at the diagnosis stage to assist the Oncologist in identifying the malignancy of a tumor .For finding accuracy of classification Linear genetic Programs, Multivariate Regression Spines (MARS), Classification and Regression Tress (CART) and Random Forests are used [7]. CART and MARS are algorithms where the number of instances required are drastically very high. Hence these methodologies cannot be deployed at situations where the number of occurances and the class instances available are of a limited number [8].

Proposed Methodology

The proposed system depicts the cancer detection & prediction of breast cancer. The user will first login to the system. If he is a new user, he will need to register with the system by filling the registration form. The new data will be added to the database. The classification algorithm will be applied to the data using Naïve Bayes & SVM algorithms & the result will be predicted. Usage of Hybrid algorithms not only provides a certain sense of improved accuracy in the proposed system but also reduces the user complexity of the system.

In the existing systems, we have noticed, several studies have been done by the experts of medicine regarding the selection of a proper dataset to train the algorithms. This has been done in collaboration with the designers of the chosen dataset.

Usage of such a database and this dataset to train improvised algorithms would be of great help to detect ductal carcinoma in its early stages.

Data set

The breast cancer data set has been taken from the UCI Repository. His dataset has five hundred and seventy six instances and ten attributes. The objective lies in to know and choose methodology consisting of techniques and algorithms best suited to induce the specified results, keeping the speed of process of the system as high as do-able at a time while not compromising with the accuracy, remodelling the planned system into absolutely economical one below multiple resource constraints. To assemble all of the selections taken in on top of directions in constructing a model that may well be used for the look, planning, implementation and achievements of project objectives optimally.

Naive Bayes

A Naive Bayes classifier is a basic probabilistic classifier in light of applying Bayes' hypothesis with solid autonomous presumption. A more illustrative term for the hidden likelihood model would be the self-deciding element display. In essential terms, a - Naive Bayes classifier expect that the nearness of a specific component of a class is disconnected to the nearness of some other element. The Naive Bayes classifier performs sensibly well regardless of the possibility that the hidden presumption is not valid.

The benefit of the Naive Bayes classifier is that it just requires a little measure of preparing information to appraise the methods and differences of the factors vital for characterization.

Numerical factors should be changed to their clear cut partners (binning) before building their recurrence tables. The other choice we have is utilizing the dissemination of the numerical variable to have a decent figure of the recurrence. For instance, one basic practice is to expect typical circulations for numerical factors.

Machine Learning

Machine learning is the study of inspiring computers to act without being explicitly customized. It is a technique for information examination that mechanizes systematic model building. Utilizing calculations that iteratively gain from information, it permits terminals to discover concealed bits of knowledge without being expressly modified where to look.

In the previous decade, machine learning has given us self-driving autos, handy discourse acknowledgment, compelling web look, and a boundlessly enhanced comprehension of the human genome. Numerous analysts think it to be the most ideal approach to gain ground towards human-level AI.

Support vector machine (SVM)

The SVM is the propelled innovation with greatest arrangement calculations installed in measurable learning hypothesis .SVM techniques are utilized as a part of characterization of straight and nonlinear information. It changes the first preparing information into higher measurement utilizing nonlinear mapping. Within this new measurement it scans for straight ideal isolating hyperplane. Information from two classes can be isolated by hyperplane with a fitting nonlinear mapping to an adequately high measurement. Utilizing bolster vectors and edges the SVM finds these hyperplane. SVM Implements the order errand by boosting the edge orders both class while limiting the grouping mistakes. In spite of the fact that the SVM can be connected to different advancement issues, for example, relapse, the great issue is that of information characterization. In machine learning, bolster vector machines (SVMs, likewise bolster vector systems) are managed learning models with related learning calculations that dissect information utilized for arrangement and relapse examination. Given an arrangement of preparing cases, each set apart as having a place with either of two classifications, a SVM preparing calculation manufactures a model that doles out new cases to one classification or the other, making it a non-probabilistic paired direct classifier. A SVM model is a portrayal of the cases as focuses in space, mapped so that the cases of the different classes are partitioned by a reasonable hole that is as wide as could be allowed. New cases are then mapped into that same space and anticipated to have a place with a class in view of which side of the crevice they fall on.

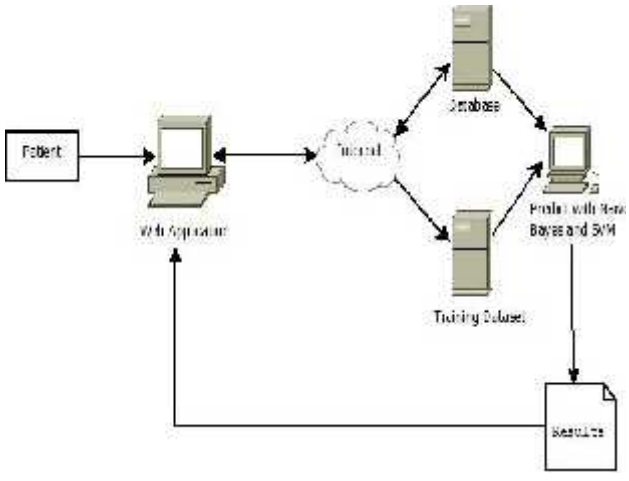


Figure 2 System Diagram for Predicting & Detecting Breast Cancer

The above diagram depicts the exact flow of the proposed system where the main deployment would be in the form of a stand-alone desktop web application, which includes the registration page, and the login page. The rest of the report shall be retrieved on the basis of the unique patient ID. Internet and PHP connectivity would be required as the primary connectivity agent of our system. This would then lead to the existing dataset which houses the training data, in order to train the algorithm on the basis of malignant or benign cases. All the data will be stored in the prepared database. Main logic of the algorithm will then be applied on the data to give the results and output of the current report. This result will be returned to the Web application and it will be displayed as the final output of our proposed system.

List of modules

User interacton

Includes the registration and login pages, where the personal details and information pertaining to the user will be taken as input and stored in our database. There will be valuation of several factors in the creation of an interactive user input page.

Pre-processing

Data pre-processing is a vital stride in the information mining process. The expression "refuse in, trash out" is especially appropriate to information mining and machine learning ventures. Information gathering strategies are frequently approximately controlled, bringing about out-of-range qualities (e.g., Pay: −100), unthinkable information mixes (e.g., Sex: Male, Pregnant: Yes), missing qualities, and so forth. Investigating information that has not been precisely screened for such issues can deliver deceiving comes about. In this way, the portrayal and nature of information is above all else before running an analysis.

In the event that there is much insignificant and repetitive data present or uproarious and untrustworthy information, then learning revelation amid the preparation stage is more troublesome. Information planning and separating steps can take significant measure of preparing time. Information pre-preparing incorporates cleaning, standardization, change, highlight extraction and determination, and so on. The result of information pre-handling is the last preparing set. Kotsiantis *et al*. (2006) introduce an outstanding calculation for each progression of information pre-processing.

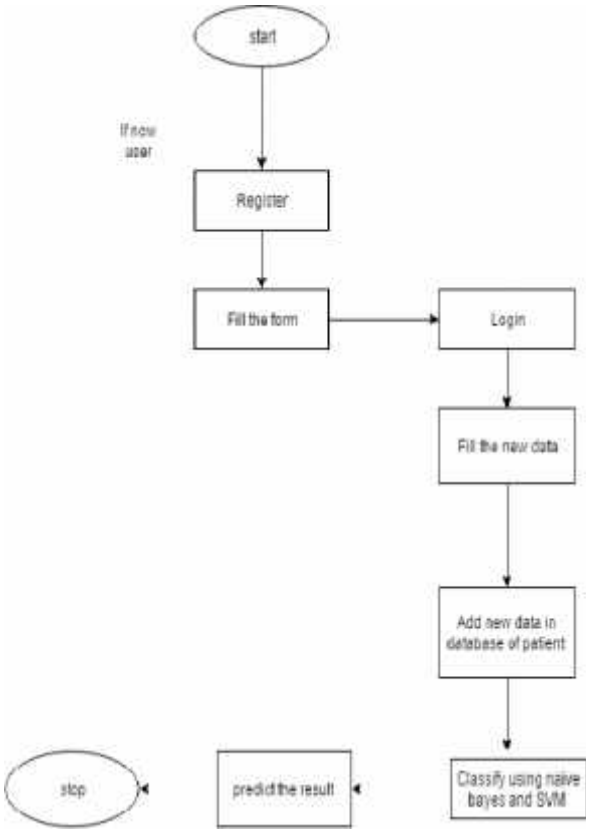
Data Storage and Training the Dataset

A database is used to load in the received data, and patient details into the server where the algorithm will be applied on the training dataset. The main purpose of using a training dataset is to indicate the benign cases prominently from the malignant ones. Data is used to model and train the dataset with certain conditions applied on data.

Report Generation and Linking It with Unique ID

Unique patient ID will be generated at the time of registration of the patient information and general details. The pre-loaded reports will be linked with the ID number generated and patient details will be retrieved. The final report will be generated and displayed to the patient.

The figure below shows the workflow of the proposed system. Prediction and detection of the final results will take place in the final stages of the working model. A dual classifier will be put to work in order to generate the most accurate and efficient results.



Applications

Applications of our system lie mainly in the healthcare and medical sectors. The system will be portable enough to run in areas where physical reach of paramedics is not possible quickly, but only after a certain period of time. Early diagnosis and detection are the keys to ensure that a patient survives ductal carcinoma. Ductal carcinoma is a form of curable malignancy which can be completely eliminated if detected early. Any institution of oncology will be able to use our system to full effect in detection of ductal carcinoma.

CONCLUSION

It is very important to diagnose and detect cancer in its early stages to ensure proper course of treatment. Breast cancer is one of the curable forms of cancer where the carcinoma is detectable quite accurately in stage 0 and stage 1. The designed system will help to increase the accuracy to its upper limits of 90%-95%. We could reduce

The complexity of this problem may be reduced by not involving machine learning techniques, but this has its own practical problems of not being able to remove outliers and noisy data from the training dataset. We propose a system which can effectively and accurately predict and diagnose breast cancer with a sample input of our parameters.

References

1. Ms. Rashmi G D Master of Computer Applications PES Institute of Technology Bangalore, India, Mrs. A Lekha Master of Computer Applications PES University Bangalore, India, Dr. Neelam Bawane Master of Computer Applications PES Institute of Technology Bangalore, India "Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset" International Conference on Emerging Research in Electronics, Computer Science and Technology – 2015
2. Alaa M. Elsayad Computers and Systems Dept, Electronics Research Institute, Cairo, Egypt. Electrical Engineering Dept, Engineering College, Salman University, Saudi Arabia, H.A. Elsalamony Mathematics Dept., Faculty of Science, Helwan University, Cairo, Egypt. Computer Science & Information Dept., Arts & Science College, Salman University, Saudi Arabia, "Diagnosis of Breast Cancer using Decision Tree Models and SVM", *International Journal of Computer Applications* (0975 – 8887) Volume 83 – No 5, December 2013.
3. Aruna Sundaram, Department of Computer Applications, Dr. M.G.R Educational and Research Institute University, Maduravoyal, Chennai, Tamil Nadu, India, Hybrid SPR algorithm to select predictive genes for effectual cancer classification, *Turkish Journal of Electrical Engineering & Computer Sciences.*
4. Bijan Moghimi-Dehkordi, Azadeh Safaee, Research Center for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Science, Tehran 1985711151, Iran, An overview of colorectal cancer survival rates and prognosis in Asia, *World J Gastrointest Oncol* 2012.
5. G.Sujatha 1Department Of Master of Computer Applications, Rao &Naidu Engineering College, Ongole, AP, INDIA, K. USHA RANI Department Of Computer Science, SPMVV(Women's University), Tirupati, ap, India, a survey on effectiveness of data mining techniques on cancer data sets, ACICE-2013 *International Journal of Engineering Sciences Research*-IJESR.
6. Ritu Chauhan "Data clustering method for Discovering clusters in spatial cancer databases" *International Journal of Computer Applications* (0975-8887) Volume 10-No.6, November 2010.
7. Dechang Chen "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering" Hindawi publishing corporation, *Journal of Biomedicine and Biotechnology* Volume 2009, Article Id 632786.
8. S M Halawani "A study of digital mammograms by using clustering algorithms" *Journal of Scientific & Industrial Research* Vol. 71, September 2012, pp. 594-600.
9. Ada and Rajneet Kaur "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient" *International Journal of Computer Science and Mobile Computing*, IJCSMC, Vol. 2, Issue. 4, April 2013, pg.1 – 6, ISSN 2320–088X
10. V.Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques"

International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646

11. Charles Edeki "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability" *Mediterranean journal of Social Sciences* Vol 3 (14) November 2012, ISSN: 2039-9340.
12. Zakaria Sulimanzubi "Improves Treatment Programs of Lung Cancer using Data Mining Techniques" *Journal of Software Engineering and Applications*, February 2014, 7, 69-77
13. Labeed K Abdulgafoor "Detection of Brain Tumor using Modified K-Means Algorithm and SVM" *International Journal of Computer Applications* (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013
14. A. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" *International Journal of Computer Science Issues*, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org
15. Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" *International Journal of Engineering, Science and Technology* Vol. 2, No. 2, 2010, pp. 59-66.

****

How to cite this article:

Kritharth Pendyala *et al*.2017, Cancer Detection & Prediction Using Dual Hybrid Algorithm. *Int J Recent Sci Res.* 8(3), pp. 15777-15789.