Visual Healthcare Analytics using Adaptive Data Mining

Priyanka B. Shivagunde
M.E. Computer Sci. & Engg
Walchand Institute of Technology, Solapur

Anita R. Kulkarni
Asst. Professor
Computer Sci. & Engg. Dept.
Walchand Institute of Technology, Solapur

ABSTRACT

In today's life health care is the very important factor. Each and every day there is new inventions coming out in health care. These inventions provide us new advanced health care services such as new techniques for detection and prevention of diseases and suggestion for new medicines. Detection of disease needs to analyze the biomedical data of human being to classify them into the fit and unfit person with respect to diagnosis of particular disease. There are different classification algorithms in data mining. Till now many researchers have used various algorithms in health care services to increase an accuracy of prediction of diseases.

This paper focuses on a proposed system used to detect disease by analyzing symptoms and test reports. It predicts whether the individual has the disease or not. It also gives the stage of the disease the person is suffering from. This system also provides the visualization of the report which helps the patients from the nonmedical background to understand the stage of the diseases. The genetic classification algorithm OlexGA is used for detection of disease and specific stage. The proposed System will work as diagnostic as well as the preventive method for individual. The an important characteristic of the proposed system is its adaptive nature of new symptoms and tests that do not exist in the system. The system is trained automatically to adapt itself to new symptoms and tests of the patients.

General Terms

Genetic Classification Algorithm OlexGA, Visualization, Adaptive data mining, healthcare

Keywords

Positive text, negative text, Symptoms, diagnostic test reports, fit person, unfit person, Carcinoma cells.

1. INTRODUCTION

Healthcare is most important factor effects on today's life. Health is most precious for human beings, Therefore, people try to take more efficient and advanced healthcare services.
Data mining is the process of extracting or mining knowledge from large data, database or any other data repositories. For analyzing data there are different algorithms are used in data mining. We can use applications of data mining in healthcare analytics. Health-related data or biomedical data is not limited, so technique of healthcare should adaptive for new symptoms
or tests mean if the text in patient's data does not exist already in a system then also the system should work properly i.e. in such case also result should correct because in healthcare applications the accuracy of the result is important. For the accuracy purpose, the data mining algorithm used for classifying fit and unfit persons is OlexGA algorithm. The main advantage of Olex GA is that it gives a most accurate result of fit and unfit person, which is required property of healthcare system.

There are different diseases, our proposed system is focused on Cancer disease. Cancer is one of the severe fetal diseases because it is not recoverable in the last stage, if it is not detected early and possibility of survivability of cancer patients is less. In women the reason death is more due to Cancer disease and most of the women from those are due to Breast Cancer. So our system is taken one case study i. e. Breast Cancer (Ca. breast).

Olex GA is the text-based classification algorithm which initially analyze the training data and classify the texts in positive texts and negative texts groups. These groups are used further to give a status of the user of the system who entered his biomedical data. If all the texts from person's biomedical data are positive then that person is fit, if not then that person is unfit further the system will detect the stage of the disease which he have currently. The proposed System will work as diagnostic as well as the preventive method for individual.

2. RELATED WORK

G. Ravi Kumar proposed the accurate classification model for breast cancer[1], in order to make full use of the invaluable information in clinical data which are ignored by most of the existing method. Different methods for breast cancer detection are explored and their accuracies are compared. Limitation of this technique is that the major information or symptoms of patients are not taken in to consideration.

K.Rajesh done difference in stages of breast cancer[2]. Algorithm is applied to SEER breast cancer dataset to classify patients into either ‘Carcinoma in Situ’ (beginning or pre-cancer stage) or ‘Malignant potential’ group. Major limitation of this technique is it take only test not symptom because in disease detection symptoms is also important factor.

Bellaachia proposed an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques [3]. The data used is the SEER Public-Use Data. He has investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5decision tree algorithms. Algorithm C4.5 is used to but these algorithms only gives different possibilities not confirm. This is the main disadvantage of this technique.

Samar Al-Qarzaie proposed the technique that used the data mining techniques and tools to find breast cancer early prediction [4]. The data mining technique and tool used was the Decision Tree technique and WEKA tool**.** This technique is not adaptive to new symptoms or new diagnostic tests.

3. SYSTEM ARCHITECTURE

3.1 System Architecture Components

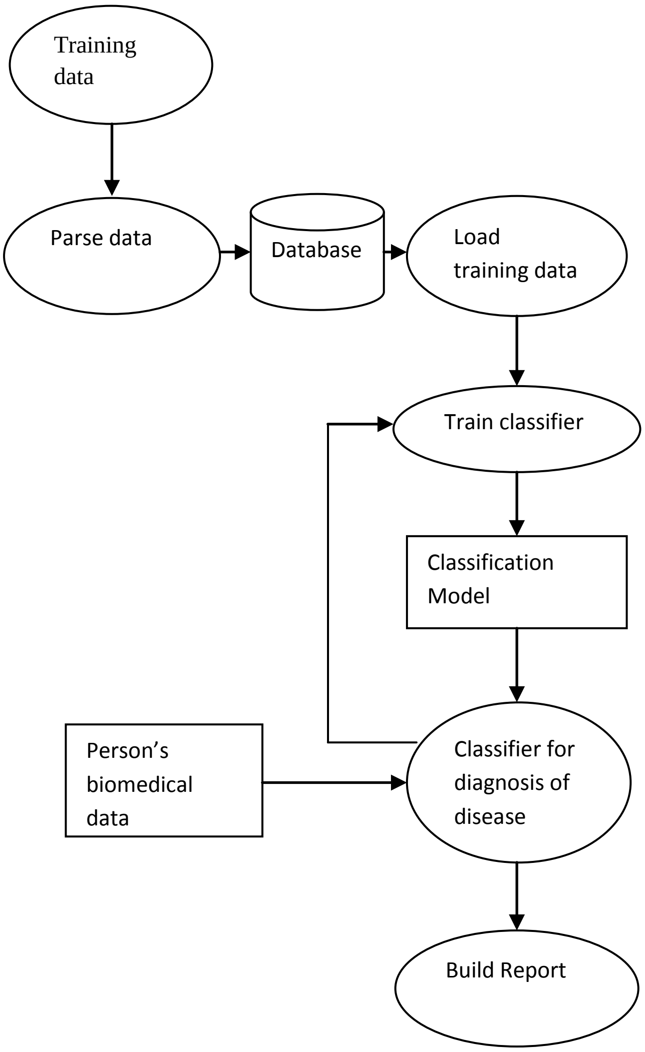


Fig 1: System Architecture

3.1.1 Training Data

The real world cancer patient's records are considered as training data. This data contains biomedical data of cancer patients including symptoms, tests, corresponding test reports, positive texts and negative texts the size of tumor etc. This record also contains the disease and stage of disease of that patients. Further, these records are analyzed and classification model will build. Records of 100 breast Cancer patients are collected from Shree Siddheshwar Cancer Hospital and Research Center, Solapur.

3.1.2 Parse Data

The training data are collected from the clinic so they are not in proper format for the analysis of data these have the data should in a proper format. To put data in proper format then have to parse data. For this parsing, the Java Parser methods are used. The parameters in the new format will be same as original format. E.g. name, Age, weight, height, symptom, test, corresponding report, a size of a tumor, disease, stage etc.

3.1.3 Train Classifier

This component used to train data, each record from training data are analyzed and trained data will get. Classify positive and negative texts: For this classification, the genetic classification algorithm Olex GA is used this algorithm takes set of all the texts from each record and identify the positive texts and negative text and after analyzing all records from loaded datasets it gives the group of

3.1.4 Classification Model

This model contains the rules of classification as fit and unfit to the user who entered his biomedical data in our system to check his fitness. The groups of positive and negative texts are present in classification model are used further to detect the disease of the new user of a system.

3.1.5 Classifier for Diagnosis of Disease

The Olex GA is used in this component to detect disease. First it take the user's biomedical data and to detect his disease it will check the texts from his biomedical data. The groups of positive and negative texts are loaded from classification model. The following method will be used to classify the patients into fit and unfit and also is current stage of disease

$C \leftarrow (t1 \epsilon d \ V \ t2 \epsilon d \ \ldots \ V \ tn \epsilon d) \ \wedge \ \neg(t_{n+1} \epsilon d \ V \ t_{n+2} \epsilon d \ \ldots \ V \ t_m \epsilon d)$

Where

c- Category

ti - term

t1…tn- positive term

tn+1…tm- negative term

d- document

3.1.6 Build Report

This component builds the final report of the user. This report contains whether the user has disease name of the disease is mentioned. A further stage of disease of a user will view by graphical representation. This graphical representation will build in this component.

3.2 Algorithm

The Olex GA Classification algorithm

1. Let a category $c \epsilon C$ and a vocabulary $V (k, f)$ over the training set *TS* to be given

2. Find two subsets of $V (k, f)$, ***Pos*** $= \{t_1, \ldots, _{in}\}$ and ***Neg*** $= \{t_{n+1}, \ldots, t_{n+m}\}$

   With $Pos \neq \emptyset$, such that H_c *(Pos, Neg)* applied to *TS* yields a ***maximum value*** of $F_{c,\alpha}$(over *TS*), for a given $\alpha \epsilon [0,1]$.

3. $C \leftarrow (T_1 \epsilon d \ V \ T_2 \epsilon d \ V \ \ldots \ V \ T_n \epsilon d) \ \wedge \ \neg(T_{n+1} \epsilon d \ V \ Tn+_2 \epsilon d \ V \ \ldots \ V \ T_{n+m} \epsilon d)$

if any of the terms $t_{1}, \ldots, _{in}$ occurs in d and none of the terms $t_{n+1}, \ldots, t_{n+m}$ occurs in d, then classify d under category c

3.3 Input Parameters

The input parameters which have to take from the new user of our system are the general information as name, age, gender, weight, height. Biomedical data as symptoms, from how many months symptoms, are noticed, diagnostic tests, corresponding reports, the size of the tumor.

3.4 Output Plan

The TNM staging system classifies cancers based on their T, N, and M stages:

- The letter T describes spread to the skin or to the chest wall under the breast.

- The letter N indicates whether cancer has spread to lymph nodes near the breast.
- The letter M indicates whether cancer has spread to distant organs -- for example, the lungs or bones.

Final result will be in graphical format: Report of safe is described in Fig 2

🟥 Presence of Carcinoma cells

🟩 Absense of Carcinoma cells

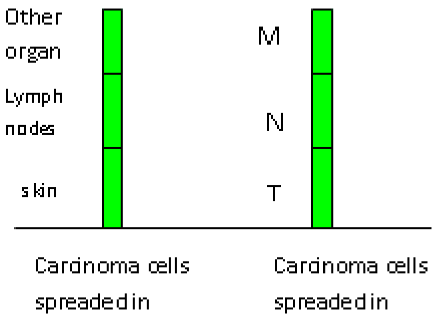


Fig 2: Safe stage

Report of patient in the first stage is described in Fig 3

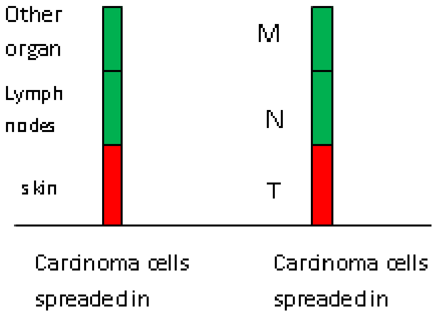


Fig 3: T Stage (First Stage of Breast Cancer)

Report of patient in the second stage is described in Fig 4

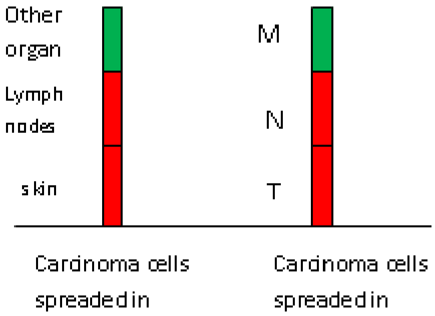


Fig 4: N Stage (Second Stage of Breast Cancer)

Report of patient in the third stage is described in Fig 5

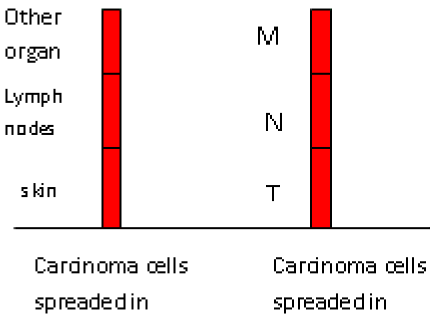
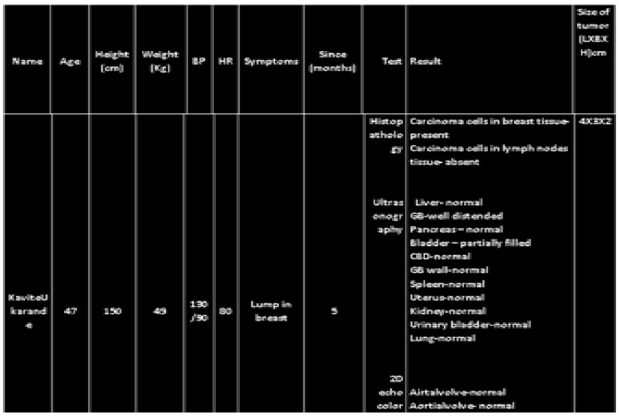


Fig 5: M Stage (Third Stage of Breast Cancer)

Example: Let user of system entered his biomedical data as in Fig 6:



Name	Age	Height (cm)	Weight (Kg)	BP	HR	Symptoms	Since (months)	Test	Result	Size of tumor (LXBX H)cm
KaviteUkarande	47	150	49	130/90	80	Lump in breast	5	Histopathology Ultrasonography 2D echo color	Carcinoma cells in breast tissue- present Carcinoma cells in lymph nodes tissue- absent Liver- normal GB-well distended Pancreas – normal Bladder – partially filled CBD-normal GB wall-normal Spleen-normal Uterus-normal Kidney-normal Urinary bladder-normal Lung-normal Airtalvolve-normal Aortialvolve- normal	4X3X2

Fig 6: Sample Data of a Patient

In above example there are carcinoma cells are found in skin cell but absent in lymph nodes and other organs than breast are normal i.e. carcinoma cells are absent in another organ so She have breast cancer and in T stage, then output will be as follows in fig 7:

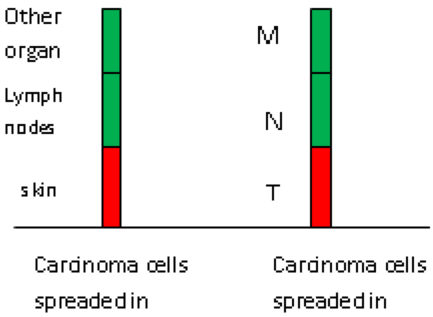


Fig 7: Report of Patient whose Historical Data Described in Fig 6

4. SCOPE

Our proposed system will work for breast Cancer disease this can be used by doctors, caretakers, friends, family or by the individual person. If there is the requirement to apply our system to any other disease than breast cancer then there have to change datasets and parameters, because framework remains same.

5. CONCLUSION

The proposed system determines that whether the user has breast cancer or not, if yes then in which stage she has currently. As our system provides data visualization of diagnosis it is easily understood by the patients having a non medical background. Our system is adaptive to new symptoms or new diagnostic tests: if medical history or data of the patient has the new symptom or new diagnostic test not existing in the system can be trained to adaptive to new symptoms without any change in the system.

6. REFERENCES

[1] "Using Data Mining Techniques for Diagnosisand Prognosis of Cancer Disease" International Journal of Computer Science,

[2] "An Efficient Prediction of Breast Cancer Datausing Data Mining Techniques" International Journal of Innovations in Engineering and Technology (IJIET)

[3] "Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm" International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012

[4] *"*Application of Data Mining Techniques to Model Breast Cancer Data*"* International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 11, November 2013)

[5] "BDCaM: Big Data for Context-aware Monitoring - A Personalized Knowledge Discovery Framework for Assisted Healthcare" IEEE Transactions on Cloud Computing, Vol. X, No. X, February 2015

[6] "An Adaptive Parameter-free Data Mining Approach for Healthcare Application" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 1, 2012

[7] "Challenges of Time-oriented Data in Visual Analytics for Healthcare"IEEE CONFERENCE PAPER · OCTOBER 2012.

[8] "A Review on Digital ECG Formats and the Relationships between Them" IEEE Transactions on Information Technology in Biomedicine, Vol. 16, No. 3, May 2012 Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[9] "A Pattern Mining Approach to Sensor-BasedHuman Activity Recognition" IEEE Transactions on Knowledge and Data Engineering, Vol. 23, NO. 9, September 2011