Diagnosis and Prognosis of Breast Cancer Using Multi Classification Algorithm

R. Shyamala
M.C.A, (MPhil)
Research Scholar,
Dept. Of Computer Science
Prist University, Tanjore
Email id: shyamganskcs@gmail.com
Ph: 7338788962

Prof. R. Maruthi
MCA, M.Phil, Ph.D
Professor,
Department of Computer Science,
Prist University
Email Id:rmaruthi2014@gmail.com

Abstract: Data mining is the process of analysing data from different views points and condensing it into useful information. There are several types of algorithms in data mining such as Classification algorithms, Regression,Segmentation algorithms, Association algorithms, Sequence analysis algorithms, etc.,. The classification algorithm can be usedto bifurcate the data set from the given data set and foretell one or more discrete variables, based on the other attributes in the dataset. The ID3 (Iterative Dichotomiser 3) algorithm is an original data set S as the root node. An unutilised attribute of the data set S calculates the entropy H(S) (or Information gain IG (A)) of the attribute. Upon its selection, the attribute should have the smallest entropy (or largest information gain) value. A genetic algorithm (GA) is aheuristic quest that imitates the process of natural selection. Genetic algorithm can easily select cancer data set, from the given data set using GA operators, such as mutation, selection, and crossover. A method existed earlier (KNN+GA) was not successful for breast cancer and primary tumor. Our method of creating new algorithm GA+ID3 easily identifies breast cancer data set from the given data set. The multi classification algorithm diagnosis and prognosis of breast cancer data set is identified by this paper.

Keywords: *Data mining, Classification algorithm, Genetic algorithm, Decision tree(ID3), medical data set.*

*****

I. Introduction

Data mining is the computational process of discovering patterns in large data sets . A method at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data mining involves six common classes of tasks, such as anomaly detection, association rule mining, clustering, classification, regression. Classification method one of the most techniques classified for large medical data set.Data mining techniques are implemented together to create a novel method to diagnosis and prognosis of breast cancer for particular patient. Genetic based ID3 algorithm is a very simplest algorithm and easily diagnosis and prognosis of cancer could be done from the given data set. Decision tree classifier does not require any domain knowledge or parameter setting. They can handle multidimensional data and are simple and past. There are many decision tree algorithms such as CART, ID3, C4.5, SLIQ, and SPRINT.

Breast cancer is considered a major health problem in men and women. In India, breast cancer cases in men and women are increasing in number. A new global study estimates that by 2030 in India increasing of breast cancer from 120,000 to around 200,000 per year. Cancer is a type of diseases which cases the cells of the body to change its characteristics and cause abnormal growth of cells. Early detection of breast cancer is essential in reducing life losses.

A prognosis is an estimate of the likely course and outcome of a disease. The prognosis of a patient diagnosed with cancer is often viewed as the chance that the disease will be treated successfully and that the patient will recover. Prognostic statements are announcements containing prognostic information. Prognostic factors are pieces of information associated with a specific outcome of disease, which can be utilized in the formulation of the prognosis.

This paper is structured as follows: section 2 the review concepts of pre processing method, Genetic algorithm,ID3 and breast cancer. Section 3 existed method. Section 4 explains our proposed method. Section 5 Results are discussed and conclusion part as section 6.

II. Basic concepts

The pre processing method using data mining techniques identify the target data from the large data set. The pre processing method has been some tasks, such as data cleaning, Data integration, Data transformation, Data reductionon, Data discretization.Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.Data integration: using multiple databases, data cubes, or files.Data transformation: normalization and aggregation.Data reduction: reducing the volume but producing the same or similar analytical results. Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called ametaheuristic) is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

Genetic algorithms are useful for search and optimization problems.GA uses genetics as its model as problem solving. Each solution in genetic algorithm is represented through chromosomes. Chromosomes are made up of genes. The collection of all chromosomes is called population. Generally three popular operators are used in GA.

1) Selection:

Selection is the stage of a genetic algorithm in which individual genomes are chosen from a population for later breeding (using the crossover operator).

A generic selection procedure may be implemented as follows:

- Fitness proportionate selection (SCX) The individual is selected on the basis of fitness. The probability of an individual to be selected increases with the fitness of the individual greater or less than its competitor's fitness.
- Boltzmann selection
- Tournament selection
- Rank selection
- Steady state selection
- Truncation selection
- Local selection

2) Crossover:

Crossover is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to the next. Cross over is a process of taking more than one parent solutions and producing a child solution from them. There are methods for selection of the chromosomes.

Various types of cross over operators are

1) Uniform crossover
2) Cycle crossover
3) Partially – mapped crossover
4) The uniform partially mapped crossover
5) Non wrapping ordered crossover
6) Ordered crossover
7) Crossover with reduced surrogate
8) Shuffle crossover

3) Mutation:

Mutation is a genetic operator used to maintain genetic diversity from one generation of a population of genetic algorithm chromosomes to the next. It is analogous to biologicalmutation. Mutation alters one or more gene values in a chromosome from its initial state. In mutation, the solution may change entirely from the previous solution. Hence GA can come to better solution by using mutation.

Fitness value:

A fitness function is a particular type of objective function that is used to summarise. Each design solution is commonly represented as a string of numbers.

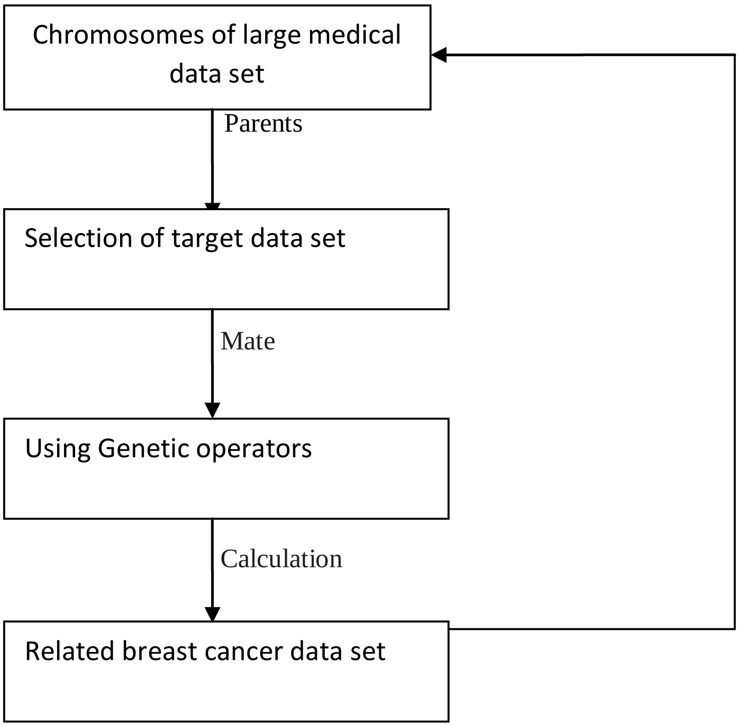
Chromosomes of large medical data set
Parents
Selection of target data set
Mate
Using Genetic operators
Calculation
Related breast cancer data set

Fig: Working genetic algorithm

III. Decision tree algorithm

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Decision trees are two types.1) classification tree 2) Regression tree. Tree models where the target variable can take a finite set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

There are many specific decision-tree algorithms:

- ID3 (Iterative Dichotomiser 3)
- C4.5 (successor of ID3)
- CART (Classification And Regression Tree)
- CHAID (CHI-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
- MARS: extends decision trees to handle numerical data better.

ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H (S) (or information gain IG (A)) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute to produce subsets of the data.

Information gain

Used by the ID3 tree-generation algorithms. Information Gain is based on the concept of Entropy from Information Theory.

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i$$

Breast cancer

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body.

Symptoms of breast cancer (female):

Breast cancer has a common medical data set:

1. Breast changes
2. Bloating
3. Between-Period Bleeding
4. Skin Changes
5. Blood in Your Pee or Stool
6. Changes in Lymph Nodes
7. Trouble swallowing
8. Weight Loss Without Trying
9. Heartburn
10. Mouth Changes
11. Fever
12. Fatigue
13. Cough
14. Pain
15. Belly Pain and depression

IV. Existed method

The existed method approach had been tested with 6 medical data sets and 1 non medical data set out of 7 data sets , 6 data sets were chosen from UCI Repository and heart disease A.P was taken from various corporate hospitals in A.P. Accuracy of the heart disease is increased by 5% using and GA using full training data set and 15% improvement in accuracy for cross validation against KNN without GA.KNN and Genetic algorithm was not successful for breast cancer and primary tumor.

V. Proposed method

Our proposed approach combines GA and Decision tree (ID3) to improve the classification accuracy of breast cancer data set. Applying Genetic algorithm for the large data set collection from medical centre using pre-processing method to identify related data set .As a result of pre-processing method and using GA operators (selection, crossover, mutation). Using GA operators we can get common attribute from medical data set. And apply Genetic results combines' decision tree algorithm identification of cancer data set. Classified the cancer data set combines of GA+ID3 and prognosis and diagnosis of breast cancer.

Genetic based ID3 classification algorithm:

```
step 1: Load the medical data set
step 2 : Apply pre-processing method on the data set  and Identify related data set
step 3: Related attribute with apply GA operators from medical data set
step 4: common data set from applying GA operators
step 5: The GA operators results  with apply ID3
step 6: apply both GA+ID3 with classified data set and getting cancer data set
step 7:  classified cancer data set with diagnosis and prognosis of breast cancer data set.
```

Accuracy of the classifier is computed as

$$\text{Accuracy} = \frac{\text{No of samples correctly classified in test data}}{\text{Total no.of samples in the test data}}$$

VI. Results and discussion

The performance our proposed method has been tested 10 data sets from medical data set and 2 non medical data set. Accuracy level increased using various data sets with genetic algorithm. The existed method using KNN with GA algorithm might be not increased breast cancer accuracy level. Our creating new algorithm GA with ID3 has increased accuracy level. This algorithm will be use to all type of cancer data sets.

VII. Conclusion

In this paper have presented classification of breast cancer using GA with ID3 algorithm. Our proposed method improving accuracy level using given the medical data set. Experiment results carried out on 10 data sets show that our approach is a competitive method for classification. The proposed method using identification cancer data set and diagnosis and prognosis of breast cancer.

References

[1] Dr.E.S.Samundeeswari (2015),"computational techniques in breast cancer diagnosis and prognosis: A Review" International journal of advanced Research ,pg.no:770-775.

[2] k.Arutchelvan, Dr.R.Periyasamy (2015)," cancer prediction system using datamining techniques" International Research Journal of Engineering and technology ,pg.no.1179-1183.

[3] Hamid Karim Khani Zand (2015)," A comoparitive survey on datamining techniques for breast cacner diagnosis and prediction",Indian journal of fundamental and applied life sciences,pg.no.4330-4339.

[4] Jaimini Majali, Rishikesh Niranjan,Vinamara Phatak,Omkar Tadakhe(2015)," data mining techniques for diagnosis and

prognosis of cancer",International journal of advanced research in computer and communication engineering.pg.no.613-616.

[5] Miss.Jahanvi joshi ,Mr.RinalDoshi,Dr.Jigar Patel,"Diagnosis and prognosis breast cancer using classification rules", International journal of engineering research and general science volume 2,pg.no.315-323.

[6] M.Akhil jabbar, B.L Deekshatulu ,Priti Chandra (2013)," Classification of Heart disease using K-Nearest Neighbor and Genetic algorithm", International conference on computational Intelligence:Modeling Techniques and applications,Elsevier ,pg.no.85-94.

[7] T.velmurugan (2014),"A survey on Breast cancer analysis using data mining techniques",IEEE international conference on computational intelligence and computing Research,pg.no.1234-1237.