International Journal of Data Mining Techniques and Applications
Volume: 06, Issue: 02, December 2017, Page No.59-61
ISSN: 2278-2419

Diagnosis and Prognosis of Oral Cancer using classification algorithm with Data Mining Techniques

N.Anitha[1], K.Jamberi[2]
[1]Assistant Professor, Department of Computer Science, K.C.S. Kasi Nadar College of Arts & Science.
[2]Assistant Professor & PhD Research scholar,
Department of Computer Science, K.C.S. Kasi Nadar College of Arts & Science, Chennai
Email: Kjamberi111@gmail.com, anithaprabhu2007@gmail.com

Abstract - Data mining is the process of researching data from different view 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 used to bifurcate the data set from the given data set and foretell one or more discrete variables, based on the other attributes in the dataset. Our method of creating new algorithm GA+ID3 easily identifies oral cancer data set from the given data set. The genetic based ID3 classification algorithm diagnosis and prognosis of oral cancer data set is identified by this paper.

Keywords: Data mining, Classification algorithm, Genetic algorithm, Decision tree, medical data set.

1. INTRODUCTION

Data mining is the computational process of discovering patterns in large data sets . It is 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 and summarization. The 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 oral 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.

Oral cancer is considered a major health problem in men and women. In India, oral cancer cases in men and women are increasing in number. A new global study estimates that by 2030 in India increase of oral cancer may be from 120,000 to around 200,000 cases per year. Cancer is a type of disease which causes the cells of the body to change its characteristics and causes abnormal growth of cells. Early detection of oral cancer is essential in reducing life losses. The estimation of the ultimate result of a disease and the analysis of the course it is likely to take is called prognosis. The hope for the successful treatment of the disease is caused by the prognosis of the affected patient. Prognostic information is generated by statement of prognosis. In the formulation of the prognosis certain bits of information are used. There pieces of information are related to the obvious yield of the disease such this of information can be called in other words prognostic factors. This paper is structured as follows: section 2: the review concepts of pre processing method, Genetic algorithm,ID3 and oral cancer. Section 3 existing method. Section 4 explains our proposed method. Section 5 Results are discussed and conclusion part is as section 6.

2. BASIC CONCEPTS

The pre-processing method using data mining techniques identify the target data from the large data set. This method have some tasks, such as Data cleaning, Data integration, Data transformation, Data reduction, Data discretization. Data cleaning: this could be defined as a process to eliminate noise and make the data coherent and consistent. By the process, values which are absent are incorporated, outliers are recognized and detected.

- Data integration: using many databases, data cubes or files.
- Data transformation: normalization and aggregation.
- Data reduction: Reducing the volume but producing the same and similar analytical results.
- Data discretization: Part of data reduction and replacing numerical attributes with nominal ones.

Natural process of selection is imitated or mimicked by GA (genetic algorithm).It is a heuristic system of search which can also be called metaheuristic. For search problems and optimization, this is used generally to find out useful solutions. The larger class of GA (Evolutionary algorithms) is born from genetic algorithms. By using techniques motivated by natural evolution, solutions are created subsequently for problems of optimization. For the purpose of identifying and solving optimization problems, genetic algorithms are widely used since they are extremely productive.GA uses genetics as its model for solution of problems. Chromosomes represent every solution in genetic algorithms. Chromosomes are made up of genes. The compiling of all chromosomes is known as population. Generally three popular operators are used in GA.

International Journal of Data Mining Techniques and Applications
Volume: 06, Issue: 02, December 2017, Page No.59-61
ISSN: 2278-2419

1) Selection

The process of isolating the Individual genomes is done by using the genetic algorithm, the same shall be used for our necessity later. This is a selection measure.

This procedure can be implemented as:

- Selection process method of Boltmann
- Tournament method of selection
- Selecting by means of Ranks
- Selection using steady states
- Selection by truncating
- Performing the selection locally (local selection)

2) Crossover

The values of chromosomes are varied from one generation to another. the process of causing such variations could be named 'crossover' more than one parent solutions are involved in crossover method. A child solution is generated from this method.

Categories of cross over operations

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

Inculcating diversity from one era of a populationto another era is called mutation. The Mutation modifies more values of the chromosomes from its existing state. While using the mutation the solution may be very distinct while comparing to the existing solution. Therefore the mutation process can be expected to give a mere good and distinct result.

Decision tree algorithm: A decision tree is a tool majorly used to support the decision making operations. It uses a tree like model while will comprise its consequences like the event of the chance outcomes, costs incurred for the resources and utilities. This method uses the decision tree like method to predict which models get them closer to the target result. Two types of decision trees exist. They are 1) classification tree 2) Regression tree. The target element in the tree model takes a finite set of values are commonly known as Classification Trees. If the target tree takes continuous values then the trees are called as Regression trees.

Algorithms for decision trees:

- 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.

We can expand ID3 as Iterative Dichotomiser 3 Mr.Ross quinlan invented this algorithm. Inorder to take a dataset and produce a decision tree from it, this ID3 is used. ID3 is used even before C4.5 algorithm . This typical algorithm enables learning through machines and is useful in the areas pertaining to the processing of natural languages. This algorithm known as ID3 starts with the root node which is the original set S. This algorithm iterates through each unutilised feature of the set S. This goes on in every iteration and leads to the calculation of the entropy H(S) of that particular attribute. The attribute with the smallest entropy or largest gain of information is chosen by it.

Information gain

Used by the ID3 tree generation algorithms. Information Gain is based on the concept of Entropy from Information Theory.inorder to generate subsets of the data . the chosen attribute is used to split the set 'S'.

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

Oral cancer

Mouth cancer is also called as oral cancer. This type of cancer makes problems in the lips,tongue,cheeks,floor of the mouth,hard and soft palate,sinuses and pharynx.

Oral cancer symptoms:

- Persistent mouth sore does not heal is symptom of oral cancer. It's a common symptom.
- Pain: Persistent mouth pain is also common oral cancer symptom
- A lump or thickening in the cheek
- A patch on the tongue and gums, either red or white in color, tonsils and mouth-lining is another symptom of oral cancer.
- A feeling that some food particle has got stuck inside the throat which cannot be washed down ,an irritation caused by this obstruction and a sore throat is also a symptom of oral cancer.
- Difficulty swallowing or chewing
- Difficulty moving the jaw or tongue
- else ware in mouth or Numbness of the tongue
- fit poorly
- Loosening of the teeth
- Pain in the teeth or jaw
- Voice changes
- Weight loss
- Persistent bad breath

3. EXISTING METHOD

The existing method approach had been tested with 7 medical data sets and two non medical data set were used to test the existing method. seven medical data sets were taken from various hospitals in Andhra Pradesh. They were chosen from UCI Repository and Heart Disease wards. Accuracy of the heart disease is increased by 5% using and GA using full training data set and 15% improvement in accuracy for cross

International Journal of Data Mining Techniques and Applications
Volume: 06, Issue: 02, December 2017, Page No.59-61
ISSN: 2278-2419

validation against KNN without GA.KNN and Genetic algorithm was not successful for oral cancer and primary tumor.

4. PROPOSED METHOD

Our proposed approach combines GA and Decision tree (ID3) to improve the classification accuracy of oral cancer data set. Applying Genetic algorithm for the large data set collection from medical centre. The pre-processing method to identify related data set and using GA operators (selection, crossover, mutation). Using these GA operators we can get common attribute from medical data set and apply Genetic results combines decision tree algorithm identification of cancer data set. This proposed method is combination of GA+ID3 using prognosis and diagnosis of oral 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 to apply with GA operators from medical data set
Step 4: common data set from applying GA operators
Step 5: The GA operator's results to apply with ID3
Step 6: Applying both GA+ID3 with classified data set and getting cancer data set
Step 7: Classified cancer data set with diagnosis and prognosis of oral cancer data set.

Accuracy of sample data has classifier computed as:

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

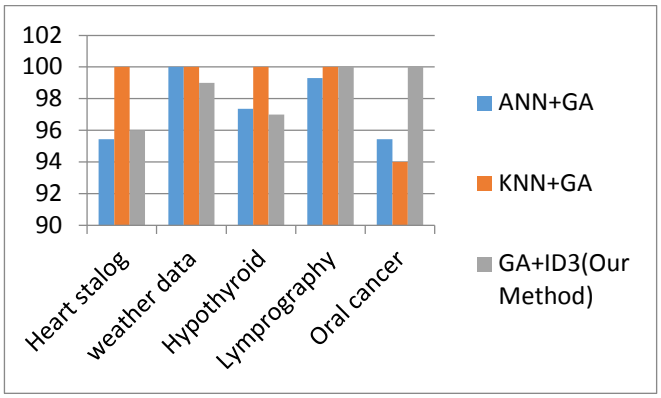
4. RESULTS AND DISCUSSION

The performance of our proposed method has been tested with 10 data sets from medical data set and 2 non medical data set. The attributes are given on the table 1 below. The Comparison of our proposed algorithm with 3 algorithms is listed in table 1.

Table 1: Accuracy comparison with various algorithms with table

Data set name	ANN+GA	KNN+GA	GA+ID3(Our Method)
Heart stalog	95.45	100	96
weather data	100	100	99
Hypothyroid	97.37	100	97
Lymprography	99.3	100	100
Oral cancer	95.45	94	100

Accuracy level increased using various data sets with genetic algorithm. The existing method using KNN apply with GA algorithm might not increase accuracy level. Our creating new algorithm GA with ID3 has increased accuracy level. This algorithm will be useful for all types of cancer data sets.



5. Conclusion

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

References

[1] 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.
[2] 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.
[3] k.Arutchelvan, Dr.R.Periyasamy (2015)," cancer prediction system using datamining techniques" International Research Journal of Engineering and technology ,pg.no.1179-1183.
[4] 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.
[5] Mr.Narayanasamy Aravindha Babu,Mr. Elumalai(2012)," Advanced Diagnostic Aids in Oral Cancer", Articles ,Research gate.
[6] 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.
[7] Dr.E.S.Samundeeswari (2015),"computational techniques in breast cancer diagnosis and prognosis: A Review" International journal of advanced Research ,pg.no:770-775.
[8] 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.