VOL. 10, NO. 9, MAY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

KNOWLEDGE EXTRACTION IN MEDICAL DATA MINING: A CASE BASED REASONING FOR GYNECOLOGICAL CANCER - AN EXPERT DIAGNOSTIC METHOD

R. Vidya[1] and G. M. Nasira[2]
[1]Manonmaniam Sundaranar University, Tirunelveli and Department of Computer Science, St. Joseph's College, Cuddalore, India
[2]Department of Computer Science, Chikkanna Government College for Women, Tirupur, India
E-Mail: vidya.sjc@gmail.com

ABSTRACT

Data mining plays an important role in Health care. It really helps to predict the disease based on collated data. Diagnosis in the medical field is a complicated task that should be performed with accuracy and efficiency. A diagnosis performed by a physician for a single patient may differ significantly if the same is examined by the other physicians or by the same physicians at different times to that single patient. Now a days, automated medical analysis are used to help doctors to predict diseases at a very fast pace. Gynecological disease is addressed in this study which is neglected; it may even prove disastrous if left unattended. Sum of 626 instances with 5 attributes are used as the medical dataset for this work. Weka is built in software tool for data mining. J48 (decision tree), Navie Bayes (NB),Multilayer perception (Artificial Neural Networks), ZeroR(Rule based) and Multiclass classifier(Support Vector Machines) are the five classification algorithms used.one of the difficult task in the predicting dermatological diseases is that as there exist a large number of similar typed presentations. This paper deals with the data collected from the southern.

Keywords: data mining, gynecology cancer, J48, navies bayes, multilayer perceptron, ZeroR, multiclass classifier, weka.

1. INTRODUCTION

Owing to the increase of various types of diseases and their definite management the health care industry faces many issues [7]. Large amount of data which is too diverse and complex to be evaluated by traditional methods are being generated by the health care transactions. The application of data mining on medical data can focus on new, useful and potentially lifesaving knowledge [8]. The extraction or mining process of knowledge from the large amount of data is said to be data mining. It is considered as an innovation which tends to help the physicians who deal with large amount of data.

The facilities that are offered by its methods ranges from interpreting complex diagnostic tests to combining information from multiple sources and there by providing support for differential diagnosis [3]. Accuracy in medical diagnostic can be improved by data mining; this reduces human resources in searching and also reduces treatment costs. Discovery of knowledge is a defined procedure in medical database [9]. Knowledge mined from data is the original concept of database.

Data can be analyzed form different views and it can be summarized into any particular idea of info [1]. Classification uses set of rules to input data into classes. It covers two types of steps, first it tries to predict the model and the function and later in the second step it goes on by applying dataset new and unseen. Based on probability the popular algorithm is Navies' Bayes. It acts as predictive model for classification task with decision tree [5]. Other algorithms based on are SVM, AI neural network, Rules based algorithm.

2. BACKGROUND

Medical database has been used widely in developed countries and it has motivated researchers in their countries to widely use data mining for knowledge discovery [2]. Day by day there derives some advancement in the database field and it aids to store lot of data. Industrialized database allows maintaining vivid patterns and permits countless knowledge extraction procedure for patient care and operative analysis of diagnosis [6].

Health organizations regularly stores and collects many data in general basics which is mostly complex. Due to this analyzing data turns out hard and decisions are made slowly. Process can be slow but patient health condition will not wait for us so it very vital to nurture a tool largely for analyzing data and for extracting information [4].

This tool can handle complex data and can derive as precise solution. Data mining when utilized more decorously a perfect task management can be done with patient data. Now the patient's details with similar type of diseases can be grouped as one and it would be easy to provide a solution to it. This method of functioning will be very effective and it would give easy way for treatment. By comparing different disease, different treatment or different type of symptoms an overall exploration can be done. Even same patients treated with different drug can be stored and analyzed and a clean and direct solution can be found out. Data mining plays a vital role in health care industry and still a large sum of different type studies can be carried out.

3. DECISION TREE

One of the most powerful tools in data mining and knowledge discovery is the decision tree. In order to discover useful patterns decision trees has been used in the examination of large complex bulk of data. The basic decision tree algorithm is called ID3 (Iterative Dichotomizer).The ID3 can handle only discrete values where as the successor C4.5 handles numeric values.

For analyzing the categorical and continuous data set Classification and Regression Trees (CART) approach is considered to be best suited. Weka developed an implementation of ID3 Algorithm called J48. Different types of data like numeric, nominal, textual data are handled by it and it also processes the missing values. Since J48 presentation is easy to understand it can be implemented in data mining packages in diverse platforms and it shows high performance with small effort.

4. NAIVE BAYES

An important role is played by the Naive Bayes classification in the medical data mining. It is a probabilistic classification based on the Bayes theorem. Due to the measure of the high input is normally very practical. All the attributes are suggested independent is called as “naïve”.

For the past 50 years machine learning method is being used in Bio-medical informatics. To estimate the parameter Naive Bayes needs only small dimensions of data set that is used for health care application. Naive Bayes The highest posterior of class instance. Simplification of assumption and naïve design, naïve Bayes classifier resolves so many real world complex problems.

Well performed Bayes classification in current approaches is:

- Boosted trees
- Random forests

5. SUPPORT VECTOR MACHINE

Linear and non-linear data can be efficiently performed by Support Vector Machine. In the year 1998, John Platt invented a method called Sequential Minimal Optimization at Microsoft Research. SMO is invented for solving optimization problem by using iterative algorithm.

By default

- It normalize all attributes
- Replaces missing values
- Transforms nominal attribute into binary ones.

6. ARTIFICIAL NEURAL NETWORKS

To solve variety of tasks the Artificial neural network is used which is very difficult to solve the problem is ordinary rules based programming that includes computer vision and speech recognition.

A feed forward artificial neural network model is called Multilayer Perceptron (MLP) which maps input data set onto a appropriate output set.

A MLP have a directed graph with multi-layered nodes in which each node has been connected to the next perspective node respectively.

Back propagation for network training on technique is used that is Supervised Learning Technique.

7. CERVICAL FACTORS

A Mostly cancer occurs due to our daily activities in our life. Growing older will bounce us lot of diseases, chewing tobacco is very dangerous, certain chemicals can also lead to cancer, certain hormones by birth will tip to cancer, Family history of cancer customarily happens a lot, Alcohol, poor diet, Lack of physical activity etc. are the foremost reason for cancer.

Smoking, chewing tobacco produces Carcinogenic agent in our body and it is more prominently points to cancer. Some of the factors analyzed are Education, Diet, Living area, Family history. More than 78 % people who are uneducated are affected than educated persons. The diet they follow unbalanced lead to 67%. Living area mostly urban people are affected then rural. Hereditary leads maximally to 80% of cancer as shown in Figure-1.

8. DATA ANALYSIS SOFTWARE

The popular machine learning software is Weka (Waikato Environment for Knowledge Analysis). Weka 3.7.9 is mainly used to analyze the data. Weka contains:

- Collection of algorithm for data analysis
- Predictive modeling
- Easy function access with GUI

Data preprocessing, classification, clustering, association rules, visualization and feature selections are standard data mining tasks which is supported by the tool Weka.

Weka has an enriched feature is

- Open source
- Platform independence

Weka provides various test options are:

- Cross validation using training set
- Test set
- Percentage split

Neural network multilayer perception performs mining and classification process by using the methods:

- Naïve Bayes
- SMO
- J48 decision trees

9. METHODOLOGY

Three stages are involved to make this research work as follows as shown in Figure-2.

First stage:

- Data collection
- Pre-processing
- Producing training data
- Analyzing variable

Second stage:

- Finding accuracy of the model by Weka tool.

Third stage:

- Prediction model explanation

10. DATA COLLECTION

Data has been collected from various place and people. We mainly targeted on the College student, working people, village people and other category of people. A complete study has been developed and data have been collected using questionnaires. After correcting missing values through filters totally we got 626 instances with 5 attributes. The five attributes are Village, Working, College, Other and Cancer. Values are counted using double and cancer through string. The cancer stands for the string B, C, O they are B for Breast cancer, C for cervical cancer and O for ovarian cancer.

11. DATA PRE-PROCESSING

Disease cases 626 database has been involved in this study. All the data sets are divided into two sections after data pre-processing.

There are 500 training sets and 100 test data set to prove the predictive model accuracy. Analyzing and implementing the data using Weka tool.

The precious information are concealed huge volume of data has been found out by data mining. Collection of machine learning algorithms for data mining is Weka tool which is written in JAVA. It includes data pre-processing, classification, regression, association rule, and clustering and visualization tools. To perform mining and classification process we deploy some methods of follows:

- Naïve Bayes method
- SMO
- J48 decision tree
- Multilayer perception

10 folds of cross validation are prune in any bias process and improve the efficiency process.

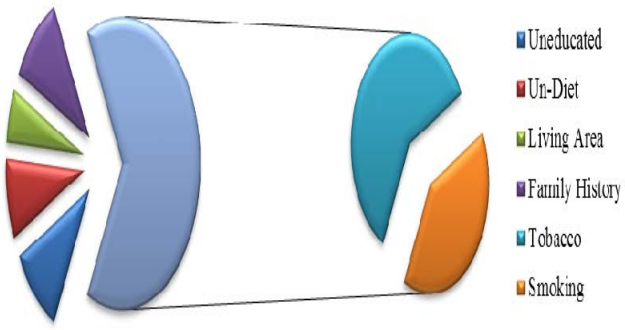

Figure-1. Foremost factors which lead to cancers.

Table-1. Weighted accuracy factor.

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC
MLP	0.986	0.009	0.984	0.986	0.985	0.978	0.994
SMO	0.912	0.066	0.887	0.912	0.891	0.852	0.898
J48	0.621	0.309	0.582	0.621	0.6	0.311	0.655
Navies Bayes	0.914	0.074	0.842	0.914	0.877	0.838	0.964

Table-2. Confusion matrix for MLP.

a	b	c	d	Classified as
0	1	0	0	a = String
0	44	3	2	b = B (Breast)
0	0	287	1	c = C (Cervical)
0	2	0	286	d = O (Ovarian)

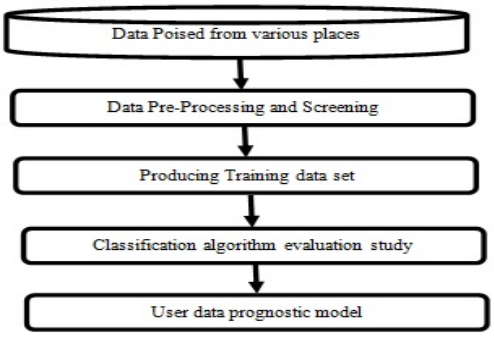

Figure-2. Structure of the work.

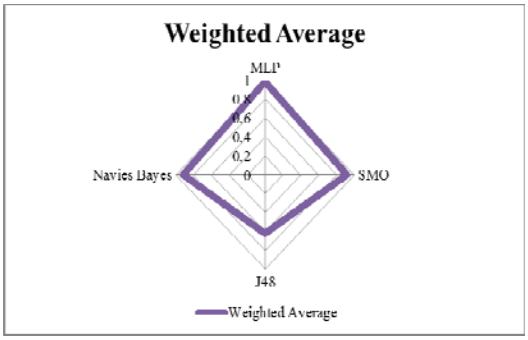

Figure-3. The weighted average of the algorithms.

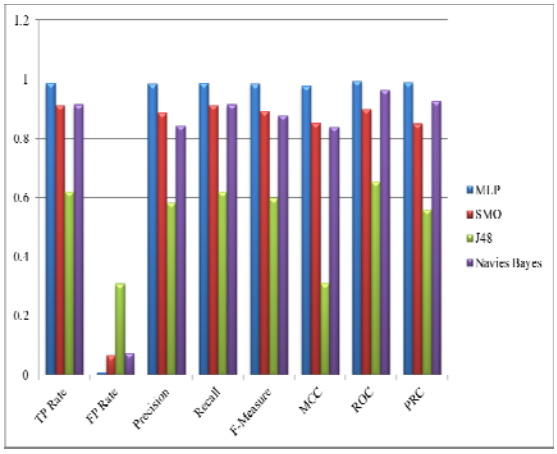

Figure-4. Result of the experiment.

Figure-5. Visualization of the result.

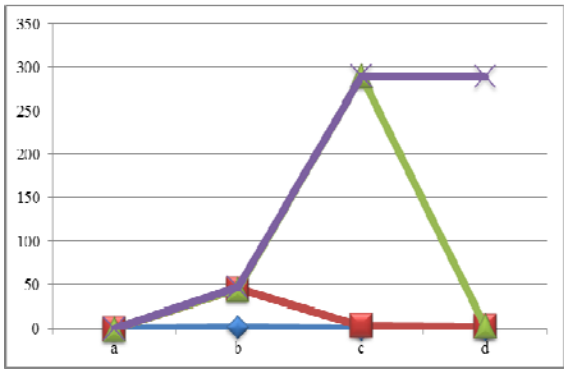

Figure-6. Graph model for confusion matrix.

12. RESULT

The total work of the algorithm is analyzed and result has been derived. Each Algorithm gives a vivid and accurate result as shown in Figure-3.

When compared multiple layer perceptron give an extreme accuracy of result (98%) but the time taken to build the model is 6.32 seconds, when next algorithm is taken into account. Support vector machine gives accuracy to (91%) and the time taken to build model is 0.21 seconds. J48 algorithm gives an accuracy of (62%) and the time taken to build model is 0.07 seconds. Navies Bayes gives accuracy to (91%) and the time taken to build model is 0.6 seconds as shown in Table-1.

The data was compared using Weka software and the result of the experiment as shown in Figure-4 and the visualization as shown in Figure-5.

When there is a excellence in a accuracy there is a drop in the time taken so when overall detailed accuracy is taken into account Navies Bayes give less performance accuracy but positive rate of speed. Multi-perceptron layer give Positive accuracy but less rate of speed. So, multilayer perceptron is more effective in the entire cross validation of accuracy class like TP-rate, FP-rate Precision, Recall, F-Measure, MCC, ROC area. The

confusion matrix by Multilayer perceptron is as shown in Table-2 along with its graph model in Figure-6.

Hence we draw a conclusion that for medical data mining Multi-Layer perceptron is really useful and it gives more accuracy than all.

13. DISCUSSION

Medical data mining is growing day by day and data are stored at rate of lakh in a second at a stretch. If in case to find a record or detail it is very hard to recollect the gynecological cancer database. To find it as fast as possible we applied four algorithms Decision Tree, Artificial Intelligence, Navies Bayes and support vector machine. The algorithm applied inside for Bayes is Navies Bayes, for support vector machine SMO, for Decision tree J48, for artificial neural network Multi-layer perceptron. MLP gives an accuracy of 98.6%. Thus the best prediction model has been identified. The instances collected are cervical cancer 290, for ovarian cancer 289 and for breast cancer 47. This can be extended to other diseases also.

REFERENCES

[1] Yang Guo, Guohuo BAi and Yan Hu. 2012. “Using Bayes Network for Prediction of Type – 2 Diabetes”, 7th International Conference for Internet Technology and Secured Transactions(ICITST), London, UK.

[2] Reza Entezarin-Maleki, Arash Rezaei and Behrouz Mimaei-Bidgoli. ”Comparison of Classification methods Based on the type of attributes and Sample Size”, Journal of Convergence Information Technology(JCIT).

[3] Boris Milovic and Milan Milovik. 2012. “Prediction and Decision Making in Healthcare Using Data Mining” Kuwait Chapter of Arabian Journal of Business and Management Review, Vol. 1, No. 12, August.

[4] Vidya R. and Dr. Nasira G.M. 2015. “A Novel Medical Support System for the Social Ecology of Cervical Cancer: A Research to Resolve the Challenges in Pap Smear Screening and Prediction at Firm Proportion”, Advances in Natural and Applied Sciences, Vol. 9, No. 6, pp. 633-638.

[5] Vidya R. and Dr. Nasira G. M. 2014. “A Theoretical study on cervical cancer Assessment & Need of Computerization in Cervical Cancer Early Detection Program”, IJETCCT, ISSN NO-2348-4454, Vol-1, pp. 109-113.

[6] Karpagavalli S., Jamuna K. S. and Vijaya M.S. 2009. “Machine Learning Approach for Preoperative Anaesthetic Risk Prediction”, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May.

[7] Vidya R., Jaia Priyankka R. P. and Nirmal Kumar G. 2014. “A system learning of connection with humans by online social networking - A rapport by means of creating usable customer Intelligence from Social media Data”, ICSEMR.

[8] Chung-Lang and Chih-Hao Chen. 2009. “Applying Decision Tree And Neural Network to Increase Quality of Dermatological Diagnosis”, Expert System with Applications, Vol. 3, pp.4045-4041.

[9] Eleni-Maria Theodoraki, Stylianos Katsaragakis and Christos Koukouvinos Christina Parpoula. 2010. “Innovative Data Mining Approaches for Outcome Prediction of Trauma Patients”, J.Biomedical Science and Engineering, Vol. 3, pp. 791-798.