Unconscious Oral Cancer Detection using Data Mining Classification Approaches

A.KALAIARASAI

Bishop Heber college

Trichy-17

Mr. K.MOHAMED AMANULLA

Asst. Professor

Bishop Heber college, Trichy-17

Abstract- **Abnormal cells divide without controlterm used for cancer diseases and able to invade other tissues.More than 100 clashing kinds of cancers are predicted in world countries. Oral cancer is an important health problem all over the world. Maximum oral cancers are recognized at a last stage where, treatment becomes fewer successful. It is very essential to identify such types of cancer at a prior stage. Prevention of oral cancer and early detection is critical, as it can growth the endurance chances substantial, allow for elementary treatment and result in an improved quality of life for survivors.In thisresearch, the datasets are acquired from various diagnostic centers which contains both cancer and non-cancer patients information and collected data is pre-processed for identical and missing information and also to proposes a three different classification algorithm are utilized that is C4.5, Random tree, and Bayesian Classification Model, Apriori algorithm and Support Vector Machine (SVM) Classification. The bestregularity for the given datasets is perform in C4.5 algorithm match up to other Classification algorithm and alsoassessment of oral cancer. Similarly it is isolated to identify the cancer and non-cancer patient's data set document. The datamining techniques will be disclosedto classify the appropriate methods and techniques for efficient classificationresults. Final contributions are the doctors in their diagnosis decisions also in their treatment process for various categories.**

Keywords: Oral Cancer, C4.5, Random Tree, Bayesian approach, Apriori Algorithm, SVM, Prediction.

1. Introduction

More deadly than breast, cervical, and prostate cancer, it has been estimated that oral cancer kills one person every hour, every day. In this studies propose that head and neck cancer and tongue cancer in specific is growing in early adults both countries. Oral cancer is a usually recognized type of head and neck cancer, which is increasing globally in occurrence and growing critically in many regions of the countries in the world. Most important step in reducing the death rate from oral cancer is early diagnosis.

Data Mining acting an important role in the Prediction of Cancer Diseases. The data mining methods judgment were aim as a main objective in many studies that mainly targeted to develop a prediction model in a dangerous fields, like medicine, by examining several data mining methods, proposing to get the model that have the highest prediction accuracy.In the classification based model the main metrics used for recognizing the presentation of every classifiers are done using were the Sensitivity, Specificity, Accurateness, Error Rate, True Positive Rate (TPR) and False Positive Rate (FPR).

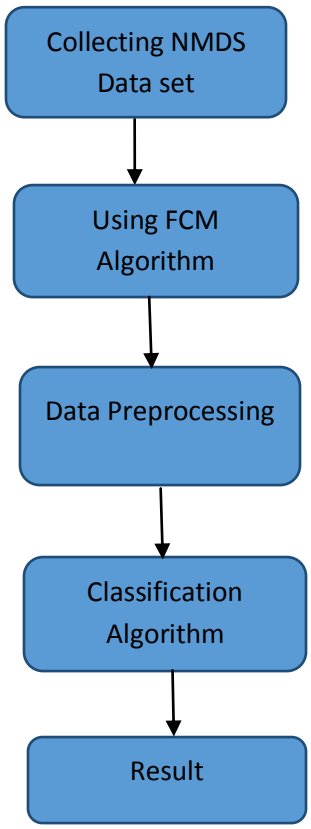
Oral cancer is a subtype of head and neck cancer and is any cancerous growth located in any sub sites of the oral cavity. It may appear as immediate lesion originating in any of the oral tissues, by metastasis

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

from a isolated site of origin, or by extension from a neighboring architectural structure, such as the nasal cavity.

The indications for an oral cancer at an earlier stage [4] are: 1) Patches inside the mouth or on lips that are white, red or mixture of white and red 2) Bleeding in the mouth 3) Difficulty or pain when swallowing 4) A lump in the neck. These indication should raise the suspicion of cancer and needs proper treatment. Therapy for Oral Cancer include surgery, radiation therapy and chemotherapy.

1.1 Data Flow Diagram



2. Literature Review

Chuang et al. [1] consider DNA repair genes. They chose single nucleotide polymorphisms (SNPs) dataset with 238 samples of oral cancer and control patients for disease prediction. All prediction analysiswere conducted using the support vector machine and they reported that the performances of the holdout cross validation was superior to 10-fold cross validation, and the best classification accuracy was 64.2%.

Kaladharet al. [3] predict oral cancer survivability using classification algorithms that include CART, Random Forest, LMT and Naïve Bayesian classification algorithms. These algorithms classify the cancer survival using 10 fold cross validation and training data set. The Random Forest classification technique correctly classifies the cancer survival dataset leading to absolute relative error relatively less as compared to other methods.

In 2010, Qin Li et al. [7] proposed to discover closed frequent item sets with a simple linear list structure called the Frequent Pattern List (FPL) in transaction database. The avenue selects representation patterns from candidate item sets to reduce combinational space of frequent patterns. By performing two procedure, signature vertex conjunction and vertex counting, it simplify the process of closed item sets generation.

In 2011, HninWintKhaing et al. [8] presented an efficient approach for the prediction of heart attack risk levels from the heart distemper database. Firstly, the heart distemper database is clustered using the K-means clustering algorithm, which will extract the data suitable to heart attack from the database. This access allows mastering the number of fragments through its k parameter. Subsequently the frequent patterns are mined from the extracted data, suitable to heart distemper, using the MAFIA (Maximal Frequent Itemset Algorithm) algorithm.

However, the oral cancer is to predict cancer and non-cancerof the every patient by using different

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

classification methods. In this research of the suggested work, the datasets are get from various diagnostic institute contains both cancer and non-cancer patient's information and collected data is pre-processes for duplicate and missing information and unpredictable of data can be utilized in different algorithm such as C 4.5, Random tree, and Bayesian Classification Model.

3. Research Methodology

3.1 Dataset

The BAHNO **NMDS (National Minimum Data Set)** is tobe the slightest quantity of data that all expert oncologist or other specialist, doctors, surgeons are managing the patients with must be expected to collect the data ofeach and every patients suffering with data set format categories are given below.

Table: Independent Variable Categories

No	Variable	Category
1	Patient ID	Numerical
2	Age	Numerical
3	Gender	Categorical
4	History of Addiction	Categorical
5	Co-Morbid Condition	Categorical
6	Symptoms	Categorical
7	Neck Node	Categorical
8	Tumor Size	Categorical
9	Cancer	Categorical

In the dataset contains number of variables included all thefields depends on the standard medical record type. Here the dataset were prepared totally 9 variables [Table] (7 input variables and 2 output variables). There is two numerical variable i.e. Case id and Age and as a Categorical variable, used Gender (Male, Female), History of Addiction (Alcohol, Smoking, Gutka, None, All), Co Morbid Condition (Hypertension, Diabetes, Immuno-compromised, none) Symptoms (No, Burning, Ulcer, Mass, Loosening of tooth), Tumor Size (<2cm, 2 cm to 4 cm, >4 cm).

3.2 Preprocessing

It containsboth cancer and non-cancer patient's information and collected data is preprocesses for duplicate and missing information and inconsistent of data and propose a classification approach is utilized.

3.3 Clustering

Fuzzy cluster means (FCM or fuzzy C-Means) model for the analysis of noisy data and clinical symptoms to categorize liver disarray. Application of cluster analysis implicatea series of procedural and diagnostic decision steps that developed the distinction and consequence of the clusters created.

3.4 Data Model

The assessment methods are mutual together using the group of data sets contains NMDS (National Minimum Data Set), will classifies the patterns in the list in order to recognize the whole database available in

http://www.aihw.gov.au/disability/disability-services-nmds-collection/

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

Clinical Symptom	Addiction	Comorbid Condition	Estimation
Burning Sensation	Tobacco Smoking	Alcohol	Plaque Like
Mass	None	None	Polypoidal
Burning Sensation	Smoking	Alcohol	Plaque-Like
Burning Sensation	Smoking	Alcohol	Plaque-Like
Mass	None	Alcohol	Polypoidal

3.5 Algorithm Specification

The Data Mining algorithm discovers the calculations, which create a data mining type from data. In this project using the efficient algorithm to utilize for a result.

3.5.1. C 4.5

The attribute values of the categories are mappedwith help of this C4.5 algorithm. The categories will beutilized for novel and unseen instances i.e. new Oral cancer records. The every row indicates a new Oral cancer record that is described through the attributes. The Iterative Dichotomizer 3 algorithm described the decision tree is a method for associated regular questioning of the attribute and their branches describe the value.

Algorithm: C 4.5

Input: An attribute at which is having valued dataset DS

1) Tree = {}

2) If DS is “empty” OR other terminate criteria met then

3) Stop

4) end if

5) for every attribute AtxDS do

6)Calculate data-theoretic criteria if we divide on attribute At

7) end for

8) Atbest= Top attribute depend upon above computed criteria

9) Tree = Construct a decision node that checks Atbest in theroot

10) DSv= Persuaded sub-datasets from DS using Atbest

11) for every DSv do

12) Treev= C 4.5(DSv)

13) Attach Treev to the equivalent branch of Tree

14) end for

15) Return Tree

3.5.2. Random Tree

It is a process to data analysis and predictivemodeling. Random tree facilitates outliers and anomaly avoidance and error detection. It succeeds best accuracy. The high dimensional data can be simply visualized with the help of random tree.

Algorithm: Random Tree Classification Algorithm

Start

{

RF= {Choose attributes subset of given dataset D}

For each chosen variable

{

If (RF.av == True) then take the relevant attributes

Else

Take the unrelevant attributes

}

for all RF until leaf node is reached.

End

Relevant attributes –cancer, Non-relevant- non-cancer

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

3.5.3. Naïve Bayes Algorithm

It is very easy to construct, not needing any complicated iterative parameter assessment schemes. This means it may be readily applied to massive data sets. It is easy to solve, so users unskilled in classifier technology can understand why it is making the classification it makes.

3.5.4.Apriori Algorithm

Apriori is a classic algorithm for learning association rules. Apriori is designed to engage on databases containing transactions. The signification of each feature is tested using KL diversity. Classification efficiency with wavelet and Gabor wavelet based texture features is also made. Wavelet family with gab or texture features leads to 92% average overall classification accuracy for preprocessing Quantization and 76.83% accuracy for Bayesian one.

Apriori Algorithm: can be used to generate all frequent Item set:

L1 = {frequent items};

For (k = 1; Lk! =∅; k++) do begin

Ck+1 = candidates generated from Lk;

For each transaction T in the databasedoes increment the

count of all candidatesin Ck+1that are contained in T

Lk+1 = candidates in Ck+1 with minimum support

end

returnCkLk;

Where, Ck: Candidate itemset of size k

Lk: frequent itemset of size k

3.5.5. Support Vector Machine (SVM)

The recurrent item sets determined by a SVM can be used to determine association rules, which hyphenate general trends in thedatabase.

Algorithm: The Candidate Generation and Test Approach.

Step 1: Initially, scan database (DB) once to get frequent 1-itemset.

Step 2: Generate length (k + 1) candidate item sets from length k frequent item sets.

Step 3: Test candidates against DB.

Step 4: Terminate, if no regular or candidate set can be produced.

4. Result and Discussion

The main aim of this paper was to determine how the data mining algorithms are consumed in the existing approaches to overcome the solution of diagnosing diseases in the earlier stages. This paper provides an idea about major life-threatening diseases and their diagnosis using classification, clustering and the bio inspirational based techniques.The Best Result producing data mining algorithms used for disease diagnosis and prognosis are shown in the figure.

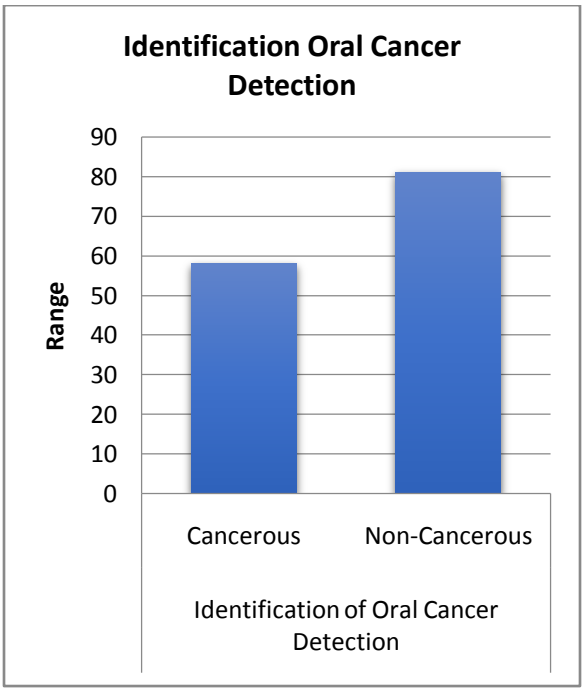


Fig 1: Oral Cancer Detection

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

Theaccuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

The accuracy and measurement of detection can be further modified by preprocessing approach. These papers focus more on accuracy. The approaches applied for cancer detection can be classified to the preprocessing difference.

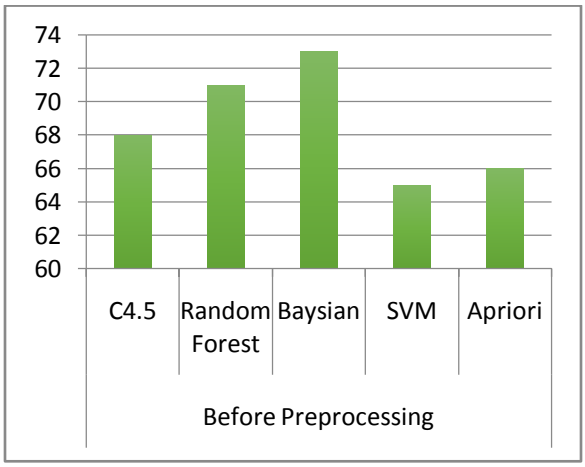


Fig 2: Before Preprocessing

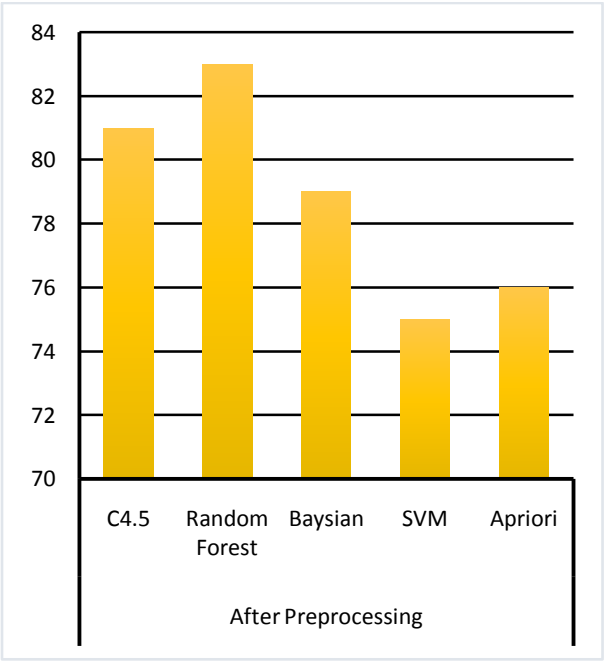


Fig 3: After Preprocessing

Evaluation of the performance of the algorithm across various demographic characteristics was conducted in the validation process.An efficient analysis technique is preprocessing to classify thedata mining techniques: Random Tree, C4.5, Naïve Bayes, SVM, and Apriori algorithms.

5. Conclusion

In healthcare, data mining is becoming increasingly more essential. In this paper, various methods to detect cancers are evaluated. The proposed work will diagnose oral cancer at an earlier stage which helps surgeons to provide medications and other treatments necessary for the particular cancer type. In this proposed technique implemented applied many classification algorithms on NMDS dataset and the performance has been analyzed five different types of algorithms. In this technique the C4.5 measure an 81%, an Apriori algorithm measure an 76%, Support vector mechanism measure an 75%, and Naïve Bayes algorithm measure an 79%. The Random forest algorithm gives a best result for detecting an oral cancer at early stage and its measures 83%. Finally the author conclude in future we combine the C4.5 and Random forest algorithm we have the better result.

6. References

[1] L. Y. Chuang, K. C. Wu, H. W. Chang and C. H. Yang, "**Support Vector Machine-based Prediction for Oral Cancer Using Four snps in DNA Repair Genes,"** Proceedings of the International MultiConference of Engineers and Computer scientists, March 16-18 2011.

[2] Kumari M. and Godara S., "**Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction**," International Journal of Computer Science and Technology (IJCST) Vol. 2, Issue 2, June 2011.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

[3] Werning, John W, Oral cancer: diagnosis, management, and rehabilitation, p. 1. ISBN 978-1588903099, May 16, 2007.

[4] Crispian Scully, Jose.V. Bagan, Colin Hopper, Joel. B. Epstein, "**Oral Cancer: Current and future diagnostics techniques – A review article**", American Journal of Dentistry, vol. 21, issue 4, pp 199 – 209, August 2008.

[5] Silverman S Jr. **Clinical diagnosis and early detection of oral cancer**. Oral MaxillofacSurgClin North Am 1993;5:199-205.

[6] DSVGK Kaladhar, B. Chandana and P. B. Kumar, "**Predicting cancer survivability using Classification algorithms**," International Journal of Research and Reviews in Computer Science (IJRRCS), vol. 02, issue 02, pp. 340- 343, April 2011.

[7] Qin Li and Sheng Chang, "**Generating Closed Frequent Itemsets with the Frequent Pattern List**", IEEE 2010.

[8] HninWintKhaing," **Data Mining based Fragmentation and Prediction of Medical Data**", IEEE 2011.

[9] A. J. Lee, W. C. Lin, and C. S. Wang, "**Mining Association rule with multi-dimensional constraints**," Journal of Systems and Software, pp: 79-92, 2006.

[10] R. T. Nguyen, V. S. Lakshman, J. Han and A. Pang, "**Exploratory Mining and Pruning Optimizations of Constrained Association Rules**," International Conference on Management of Data, ACM-SIG-MOD. 13-24, 1998.

[11] B. Milovic and M. Milovic, "**Prediction and decision making in health care using data mining**". International Journal of Public Health Science. 01, 02 (December 2012), 69-78, 2012.

[12] K. Anuradha, and K. Sankaranarayanan, "**Identification of suspicious regions to detect Oral cancers at an earlier stage– a literature Survey**," International Journal of Advances in Engineering & Technology, vol. 03, issue 01, pp: 84-91, March 2012.

[13] J. Nahar, S. T. Kevin, A. B. M. S. Ali and Y. P. Chen, "**Significant cancer prevention factor extraction: An association rule discovery approach**," J Med Syst, Springer, October 2009, DOI 10.1007/s10916-009-9372-8

[14]A.Sudha "**Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability**", March 2012.

[15] K.Balachandran, Dr. R.Anitha "**Supervised Learning Processing Techniques for Pre-Diagnosis of Lung Cancer Disease**", 2010.

[16]Hemantpalivelasurve .et.al. "**On Mining Techniques for Breast Cancer Related Data**", 2012.

[17] Arihito Endo et. al., "**Comparison of Seven Algorithms to Predict Breast Cancer Survival Biomedical**", 2008.

[18] G. Cong, and B. Liu, "**Speed-up IterativeFrequentItemset Mining with Constraint Changes**," ICDM, pp: 107- 114, 2002.

[19] K. RuthRamya et al, "**A Class Based Approachfor Medical Classification of Chest Pain**", International Journal of Engineering Trends and Technology, Vol. 3, Issue. 2, pp.89-93, 2012.

[20] S .Swami et al, "**MultidimentionalAssociationRules Extraction in Smoking Habit Database**", International Journal of Advanced Networking and Applications, vol: 03, issue: 03, pp. 1176- 1179 (2011).

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

[21] Sung Ho Ha and SeongHyeonJoo, "**A HybridData Mining Method for Medical Classification of Chest Pain**", World Academy of Science, Engineering and Technology, 37, pp: 608-613, 2010.

[22] Rui Chang; Zhiyi Liu, "**An improved apriorialgorithm**," Electronics and Optoelectronics (ICEOE), 2011 International Conference on, vol.1, no., pp.V1-476, V1-478, 29-31 July 2011.

[23] Jitao Zhao and Ting Wang, "**A General Framework for Medical Data Mining**", 2010 International Conference on Future Information Technology and Management Engineering.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET