IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

A New Feature Selection Method for Oral Cancer Using Data Mining Techniques

Mrs. R. Vidhu[1], Mrs. S. Kiruthika[2]

Assistant Professor, Department of BCA, Vidyasagar College Of Arts and Science, Udumalpet, India [1]

Assistant Professor, Department of Computer Science, P.K.R Arts College for Women, Gobi, India [2]

Abstract: The term cancer is used generically for more than 100 different diseases including malignant tumours of different sites (such as breast, cervix, prostate, stomach, colon/rectum, lung, mouth, leukaemia, sarcoma of bone, Hodgkin disease, and non-Hodgkin lymphoma). Common to all forms of the disease is the failure of the mechanisms that regulate normal cell growth, proliferation and cell death. Ultimately, there is progression of the resulting tumour from mild to severe abnormality, with invasion of neighbouring tissues and, eventually, spread to other areas of the body. The primary risk factor for developing oral cancer is tobacco use. Smoking cigarettes, cigars, and pipes all increase risk of oral cancer. Smokeless tobacco, often called "dip" or "chew," also heighten the risk. Alcohol consumption is another habit that is strongly associated with the development of oral cancer. This research uses data mining technology such as classification, clustering and prediction to identify potential oral cancer patients.Apriori algorithm is the originality algorithm of Boolean association rules of mining frequent item sets. The datamining methods and techniques will be explored to identify the suitable methods and techniques for efficient classificationof data.The data mining techniques are effectively used to extract meaningful relationships fromthese data. Genetic algorithms were applied to association and classification techniques.

Keywords: Cancer; Oral Cancer; Data Mining; Classification; Decision tree; Apriori algorithm; Association rule; Genetic algorithm.

1. INTRODUCTION TO ORAL CANCER

Good dental or oral care is important to maintaining healthy teeth, gums and tongue. Oral problems, including bad breath, dry mouth, canker or cold sores, TMD, tooth decay, or thrush are all treatable with proper diagnosis and care. Oral cancer can affect any area of the oral cavity including the lips, gum tissues, tongue, cheek lining and the hard and soft palate.

Possible signs and symptoms of oral cancer when patients may report include: a lump or thickening in the oral soft tissues, soreness or a feeling that something is caught in the throat, difficulty chewingor swallowing, ear pain, difficulty moving the jaw or tongue, hoarseness, numbness of the tongue or other areas of the mouth, or swelling of the jaw that causes dentures to fit poorly or become Uncomfortable. Other symptoms of oral cancer may include:

- a sore or blister in your mouth or on your lip that does not heal after two weeks
- lesion on the tongue or tonsil
- white and red patches in the mouth or lips that does not heal
- bleeding from the mouth that is unrelated to an injury
- change in the way teeth fit together, including how dentures fit or loose teeth because of jaw swelling or pain
- persistent earaches
- difficulty swallowing, chewing, speaking, or moving the tongue.

2. WHAT ARE THE RISK FACTORS FOR ORAL CANCERS?

- **Tobacco**

All forms of tobacco increases a person's risk of oral cancer. In fact, smokers are six times more likely to get an oral cancer than non-smokers.

DOI 10.17148/IJARCCE.2016.5108

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

- **Alcohol**

Heavy, regular alcohol consumption is a risk factor for oral cancer. It's estimated that 75 to 80 percent of those with oral cancer drink alcohol frequently. Like smokers, people who drink a lot of alcohol on a regular basis are also six times more likely to get an oral cancer than non-drinkers.

- **Gender**

It appears that men contract oral cancer at twice the rate of women, due to the fact that they are more likely to smoke and drink heavily for longer periods of time than females.

- **Age**

After the age of 40, the risk of oral cancer increases, with 60 being the average age of diagnosis.

Other Risk Factors

- Viral infections
- Immuno-deficiencies
- Poor nutrition
- Exposure to ultraviolet light (responsible for many cases of cancer to the lips)
- Certain occupational exposures

Signs and Symptoms of Oral Cancer

- Sores in the mouth or on the lips that don't heal and/or bleed easily.
- A white or red patch of skin in the mouth or under the tongue that doesn't go away.
- A lump in the mouth, throat, or tongue.
- A sore throat that doesn't go away within a normal period of time.
- Swallowing and/or chewing is difficult or painful

Screening for oral cancer

Early detection (as distinct from organized screening) of oral cancer using visual inspection of the mouth is being considered in countries where incidence is high, such as Bangladesh, India, Pakistan, and Sri Lanka. The oral cavity is easily accessible for routine examination, and nonmedical personnel can readily detect lesions that are the precursors of carcinoma (WHO, 1984). Furthermore, there are indications that precursor lesions may regress if tobacco use ceases, and that surgical treatment of early oral cancer is very effective. Experience in south-east Asia has demonstrated under field conditions that primary health care workers can examine large numbers of people, and detect and classify precancerous lesions and cancers of the oral region with acceptable accuracy. Some programmes have also encouraged early detection of oral cancer by self-examination using a mirror (Mathew et al., 1995).However, so far it has not been shown that detection of precancerous lesions or early cancers can reduce mortality from the disease.

Association Rule Mining Technique in oral cancer

What is association Rule Mining?

Data mining technique, association rule mining is applied to search the hidden relationships among the attributes. It identifies strong rules discovered in databases using different measures of interestingness. Thus, an association rule is a pattern that states when X occurs, Y occurs with certain probability. Association rule mining proceeds on two main steps. The first step is to find all item sets with adequate supports and the second step is to generate association rules by combining these frequent or large item-sets. In the traditional association rules mining, minimum support threshold and minimum confidence threshold values are assumed to be available for mining frequent item sets, which is hard to be set without specific knowledge; users have difficulties in setting the support threshold to obtain their required results. To use association rule mining without support threshold, another constraint such as similarity or confidence pruning is usually introduced.

Association Rule Mining is all about finding all rules whose support and confidence exceed the threshold, minimum support and minimum confidence values. In the traditional association rules mining with FPtrees and reduction technique, minimum support threshold and minimum confidence threshold values are assumed to be available for mining frequent item sets, which is hard to be set without specific knowledge; users have difficulties in setting the support threshold to obtain their required results. Setting the support threshold too large, would produce only a small number of rules or even no rules to conclude. In that case, a smaller threshold value should be guessed (imposed) to do the mining again, which may or may not give a better result, as by setting the threshold too small, too many results would be produced for the users, too many results would require not only very long time for computation but also for screening these rules.

Soft computing techniques

Soft Computing refers to a collection of computational techniques in computer science, artificial intelligence, machine learning and some engineering disciplines, which attempt to study, model and analyze very complex phenomena. Earlier computational approaches could model and precisely analyze only relatively simple systems. More complex systems arising in biology, medicine, the humanities, management sciences, and similar fields often remained intractable to conventional mathematical and analytical methods. This is where soft computing provides the solution.

Key areas of soft computing include the following:

1. Fuzzy logic

Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic. It can be thought of as the application side of fuzzy set theory dealing with well thought out real world expert values for a complex problem. Fuzzy logic can be used to control household appliances such as washing machines (which sense load size and detergent concentration and adjust their wash cycles accordingly) and refrigerators.

2. Neural Networks

Artificial neural networks are computer models built to emulate the human pattern recognition function through a similar parallel processing structure of multiple inputs. A neural network consists of a set of fundamental processing elements (also called neurons) that are distributed in a few

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

hierarchical layers. Most neural networks contain three types of layers: input, hidden, and output. After each neuron in a hidden layer receives the inputs from all of the neurons in a layer ahead of it(typically an input layer), the values are added through applied weights and converted to an output value by an activation function (e.g., the Sigmoid function). Then, the output is passed to all of the neurons in the next layer, providing a feed forward path to the output layer.

3. Statistical Inferences

Statistics provides a solid theoretical foundation for the problem of data analysis. Throughhypothesis validation and/or exploratory data analysis, statistical techniques give asymptoticresults that can be used to describe the likelihood in large samples. The basic statistical exploratory methods include such techniques as examining distribution of variables, reviewing large correlation matrices for coefficients that meet certain thresholds, and examining multidimensional frequency tables.

4. Rule induction

Induction models belong to the logical, pattern distillation based approaches of data mining.

Based on data sets, these techniques produce a set of if-then rules to represent significant patterns and create prediction models. Such models are fully transparent and provide complete explanations of their predictions**.**

5. Genetic Algorithm

They are search algorithms based on the mechanics of natural genetics i.e. operations existing in nature. They combine a Darwinian 'survival of the fittest' approach with a structured, yet randomized, information exchange. The advantage is that they can search complex and large amount of spaces efficiently and locate near optimal solutions rapidly.

The algorithm operates through a simple cycle as shown in fig .

1. Creation of a population of strings
2. Evaluation of each string
3. Selection of the best strings
4. Genetic manipulation to create a new population of strings.

The GA maps strings of numbers to each potential solution. Each solution becomes an individual in the population and each string becomes a representation of an individual.

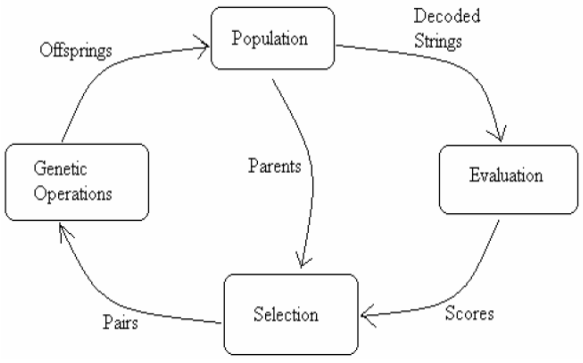

Figure1: The Genetic Algorithm cycle

Two operators are used: Crossover and Mutation. The offspring generated by this process replaces the older population and the cycle is repeated until a desired level of fitness is achieved .Crossover is one of the genetic operators used to recombine the population's genetic material. It takes two chromosomes and swaps part of their genetic information to produce new chromosomes. The mutation operator introduces new genetic structures in the population by randomly changing some of its building blocks. Genetic algorithms were applied on the data mining techniques to improve their search procedure and reduce programming complexity. GA was applied on the association and classification rule mining techniques.

Application of Genetic algorithm into Association rule mining

GA was applied to association rule mining in the following manner

Algorithm Genetic

Input: Database that was taken for association rule mining
Read the initially provided random rules that are of the form '100100100' which implies out of 9 items, item 1, item 4 and item 7 occur together in the database.do while consecutive generations are equal
call function Means() // to determine the meaning of the representation.
call function Fit() //determine fitness value i.e. number of times the rule
//occurs in the database
call function Reproduction() // select most fit rules and make their copies
call function Crossover() // cross over the rules to create new rules
call function Means ()
call function fit()
end while

Apriori Algorithm

The apriori is a classic algorithm for frequent item set mining and association rule learning over the transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by a apriori can be used to determine association rules, which highlight general trends in the database.

Association rules mining using apriori algorithm uses a "bottom up" approach, breadth-first search and a hash tree structure to count the candidate item sets efficiently. A two-step apriori algorithm is explained with the help of flowchart as shown in **Figure 2**and the algorithm is mentioned below:

Apriori algorithm: 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 frequent or candidate set can be

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Generated.

To select interesting rules from the set of all possible rules generated, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

Support: The rule holds with support supp in T (the transaction data set) if supp% of transactions contain X□Y Supp(X→Y) = P(X □Y) .

Confidence: The rule holds with confidence confin T if conf% of transactions that contain X also contain Y

Conf(X →Y) = P(Y X) = Supp(X □Y) Supp(X) = P(X and Y) P(X)

Lift: It is the probability of the observed support to that expected, if X and Y were independent

Lift (X →Y) = Supp(X □Y) Supp(X)× Supp(Y) = P(X and Y) P(X) P(Y)

Leverage: It measures the difference of X and Y appearing together in the dataset and what would be expected if X and Y were statistically dependent

Lev (X→Y) = P(X and Y) − (P(X) P(Y))

Conviction: It is the probability of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction)

Conv(X →Y) = 1− Supp(Y) 1− conf(X →Y) = P(X) P(not V) P(X and not Y)

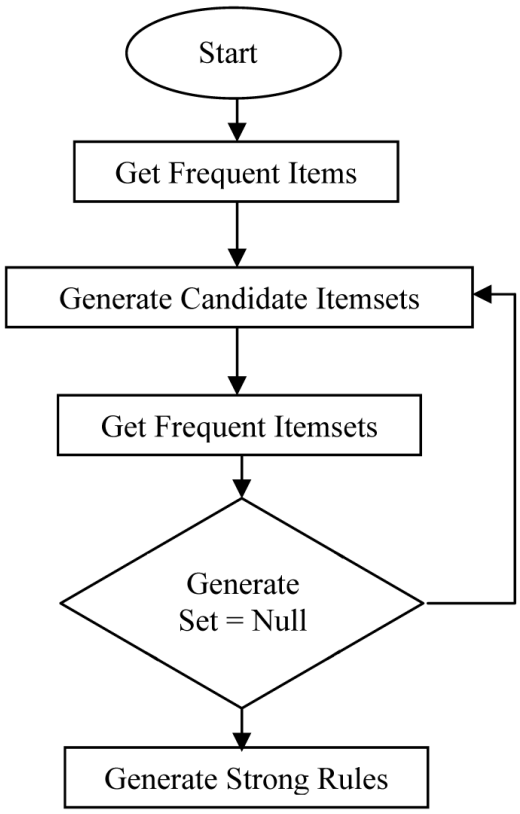

Figure 2: Flowchart of Apriori algorithm

3. CONCLUSION AND FUTURE WORK

Data mining is the non-trivial extraction of implicit previously unknown and potentially usefulinformation about data. Here, we use data mining techniques in oral cancer treatment. Data mining techniques have beenwidely used for oral cancer diagnosis. In this paper we have discuss some of effectivetechniques that can be used for oral cancer classification .In this report an in-depth study of the varied data mining techniques was made. It was shown

how genetic algorithms can be used to optimize the data mining algorithms.We have used classification, clustering algorithmic methodsand soft computing techniques for better prediction and understand ability of oral cancer in an earlier stage. Data mining is therefore an effective technique to solve the problem of enormous data faced by researchers, In the futureInnovation in diagnostic features of tumours may play a central role in development of efficient treatment methods for Oral cancer affected patients. Also shall involve applying data mining technique to diagnose and identifying the stages, and treatments of oral cancer. In future better techniques can be incorporated with the present research work for less complexity and better learning adaptability. Moreover, better fuzzy techniques could also be used to improve the classification rate and accuracy.

REFERENCES

[1] K. Anuradha, Dr. K. Sankaranarayanan“ International Journal of Advanced Research in Computer Science and Software Engineering”Volume 5, Issue 1, January 2015.

[2] International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013

[3] Neha Sharma, Hari Om-” Extracting Significant Patterns for Oral Cancer Detection Using Apriori Algorithm”

[4] S. Warnakulasuriya, "Global epidemiology of oral and oropharyngealcancer”, April-May 2009.

[5] Hipp, Jochen ,Guntzer, Ullrich and Nakhaeizadeh, Gholamreza, “Algorithms for Association Rule Mining – A general Survey and Comparison”. SIGKDD explorations, Vol 2, Issue – 1, pp 58 – 63, Mar – 2004.

[6] Nikhil SureshkumarGadewal, Surekha Mahesh Zingde, “Database andinteraction network of genes involved in oral cancer”, (2011).

[7] World Health Organization (2002) Department of Management of Noncommunicable Diseases: National Cancer ControlProgrammes, World Health Organization, Geneva.

[8] Kaladhar, D.S.V.G.K., Chandana, B. and Kumar, P.B. (2011) Predicting Cancer Survivability Using ClassificationAlgorithms. International Journal of Research and Reviews in Computer Science (IJRRCS), 2, 340-343.

[9] Sharma, N. and Om, H. (2012) Framework for Early Detection and Prevention of Oral Cancer Using Data Mining. International Journal of Advances in Engineering & Technology, 4, 302-310.

[10] Hemantpalivelasurve .et.al. “On Mining Techniques for Breast CancerRelated Data”, 2012

[11] Mallika and Saravanan, "An SVM based Classification Method for Cancer Data using Minimum Microarray Gene Expressions", World Academy of Science, Engineering and Technology, Vol. 62, No. 99, pp.543-547, 2010. www.cancer.org/clinicaltrials. Viewed 28-August- 2013. http://www.scribd.com/doc/28249613/Data-Mining-Tutorial

DOI 10.17148/IJARCCE.2016.5108