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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

Oral Cancer Detection Using Apriori Algorithm

R.J. NOEL BLESSY[1], K.MOHAMED AMANULLAH[2]

M.Phil, Scholar, Department of Computer Applications, Bishop Heber College (Autonomous), Trichy, India[1]

Assistant Professor , Department of Computer Applications, Bishop Heber College (Autonomous), Trichy, India[2]

Abstract: Oral cancer is among the 10 most common cancers in worldwide, and is especially seen in disadvantaged elderly males. Early prevention and detection and of oral cancer is critical, as it can increase the existence chances significantly, allow for simpler action and result in a better quality of life for survivors. In this article, the association rule mining algorithm, apriori is used to find the spread of oral cancer with the help of various inquiries and then measure the chance of survival of the patient. This is attained by extracting a set of rules among various laboratory tests and inquiries like RFT, fine-needle aspiration cytology (FNAC), Biopsy, Ultrasound, Sonography Test (USG), Chest x-ray or computed tomographyand survivability of the oral cancer patients. The rules clearly show that if FNAC, USG and Chest x-ray or computed tomography, then chance of existence is reduced. However, if RFT is now normal, the probability of survival is positive. If diagnostic-biopsy results in squamous-cell-carcinoma, then it clearly designate oral cancer, which may lead to high humanity if appropriate treatment is not initiated. The new results demonstrate that all the generated rules hold the highest confidence level, thereby, creation investigations very essential to understand the spread of cancer after medical examination for early detection and prevention of oral cancer.

Key-Words: Data Mining, Association Rule Mining, Apriori, Oral Cancer

I. INTRODUCTION

Knowledge discovery is the "non-trivial process of identifying valid, novel, possiblyvaluable and ultimately understandable patterns in data" [1]. Data mining steps in the procedure of knowledge discoverycontains of applying data analysis and discovery algorithms that produce a particular enumeration of our models over the data [2]. Data mining could also be characterized as the process of finding useful patterns or meaning in the raw data which subsequently can be used to develop a predictive model [3]. It has variously been called as knowledge discovery in databases(KDD), knowledge extraction, knowledge discovery, information discovery, data archeology, information harvesting and data pattern processing [4].

Knowledge discovery contains the additional steps of target data set selection, and data reduction, data preprocessing which occur prior to data mining. This is also involves the additional steps of interpretation, informationand consolidation of the information extracted during the data mining process. These extracted patterns will delivervaluable knowledge to decision makers. The data mining Application are manyit has been utilized seriously and also widely by dealers, for direct marketing and up-selling or cross-selling; by financial organizations, for credit counting and fraud detection; by manufacturers, for quality control and maintenance arrangement and by sellers, for market division and store layout. Data mining is becoming progressively popular, if not increasingly essential in the healthcare industry as well. It is so because current medicine and many healthcare transactions generate almost every day, huge amount of heterogeneous data that is too complex and voluminous to be processed and analyzed by traditional methods.

Those who deal with such data understand that there is a widening gap between data comprehension and data collection. Computerized systems are needed to help humans address this problem [5]. Data mining provides the technology and methodology to transform these mounds of data into useful info for decision creation, with the meaning of offer valuable quality services at sensible costs, which is a main concern,imagine by the healthcare organizations that is hospitals and medical centers. The Data mining applications can incredibly profit all stakeholders of the healthcare industry such as clinics, hospitals, physicians, and patients, for example, by identifying the best treatments and best performs. Data miningdiscovers hidden information in the data, however, itcannot tell the world of the info to yourinstitute. Important patterns might already beknown as a result of acquaintance to the business domain and working with data over time. However, data mining can confirm or qualify such empiricalobservations in addition to finding new patterns thatmay not be immediately discernible through simpleobservation.

Here undertaken the study of oral cancer, which is of significant public health importance in India. Public health officials, hospitals, and academic medical centres within India have recognized oral cancer as a grave problem [6]. Possible signs and symptoms of oral cancer when a patient may report include: a lump or thickening in the oral soft tissues, soreness or a touch that rather is caught in the throat, difficulty eating or swallowing, ear pain, difficulty moving the tongue or jaw, hoarseness, numbness of the tongue or other arenas of the oral cavity, or swelling of the jaw that causes dentures to fit poorly or get uncomfortable. The above mentioned signs and symptoms observed in the patients on their visit to OPD must be used efficiently for early detection and treatment, as they are critical and can increase the survival chances considerably, permit for simpler handling and effect in a fuller quality of life for survivors. In the earlier paper "Significant Pattern for Oral Cancer Detection: An Association Rule Mining" written

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

by the authors, many significant roles among various valuable information pertaining to clinical inspection, past and survivability of the cancer patients were extracted to assist the practitioners in early detection of the disease and prediction of distribution of cancer in the oral cavity, and consequently help in prevention of the oral cancer. After successfully generating vital rules from the details of clinical examination, here now extend our work by attempting to find the spread of cancer with the help of various investigations and then assess the chance of survival of the patient. In this research paper, the popular association rule mining algorithm, apriori is used to extract a set of significant roles among various laboratory tests and investigations like FNAC of neck node, Biopsy, LFT, USG, CT scan-MRI and survivability of the oral cancer patients.

II. ORAL CANCER

Oral distortion is a heterogeneous assembly of tumors rolling out from diverse parts of the oral cavity, with typical predisposing issues, prevalence, and treatment outcomes. Oral cancer is one of the ten most incessant diseases in worldwide with a yearly incidence of over 400,000 cases, of which 64% rise in developing nations [7]. There is a huge contrast in the rate of oral cancer in diverse regions of the worlds. The age-adjusted rates of oral cancer differ from over 20 for every 100,000 people in India, to 10 for every 100,000 people in the U.S., and less than 2 for every 100,000 people in the Middle East. In comparison with the U.S. people, where the oral cavity malignancy represents only about of 3% of malignancies, it accounts for over 30% of all developments in India. It has been estimated that 83,000 new oral tumor cases occur every year in India. The difference in incidence and pattern of oral tumor is due to regional changes in the prevalence of risk factors. But as the oral tumorhas well-defined risk factors, these may be better, giving real confidence for maindeterrence.

The clinician's problematic is separating malignant lesions from a nearly unlimited amount of other poorly characterized, questionable, and crudely understood lesions that also occur in the oral cavity. The oral injuries are benign, yet many have a manifestation that may be efficiently befuddled with threatening lesions and some are now considered pre-malignant because they have been statistically correlated with subsequent cancerous modifications [8]. On the other hand, some malignant lesions seen in an initial stage may mistake for a benign. Initial carcinomas are presumably asymptotic and ensuing signs are frequently misjudged in light of the fact that they imitate numerous benevolent lesions and the distress is insignificant. Professional consultation is thus often delayed,growing the chance for local spread and regional metastases. Stress must be placed on fast access tohigh risk individuals for periodic oral examinations and educational efforts to grow the skill of primary healthcare providers in recognizing this problematic. Squamous cell carcinoma accounts for 90% of the total number ofmalignant oral lesions. Therefore, the difficulty of oral cancer is primarily that of diagnosis, pathogenesis, andmanagement of squamous cell carcinoma originating from the oral muscular surface. The aim of this work is toapply to the association rule mining on the data pertaining to medical symptoms, history of addiction, co-morbid conditionand survivability in order to evaluate the medical features, diagnosis, and treatment of oral cancer patients. Various laboratory tests and investigating techniques that can be used for assessing the extent of oral cancer in the patient's body are as follows:

2.1 Renal Function Tests (RFT), Analysis of urine and blood samples can be essential for the evaluation of body kidney function. Some of the basic function tests are:

Blood urea nitrogen (BUN) provides anunevencapacity of the glomerular filtration rate, the filtration rate at which blood is filtered in the kidneys. In the body urea is formed in the liver as an end product of protein metabolism and is accepted to the kidneys for defecation. Nearly all kidney infections cause inadequate defecation of urea, elevating BUN levels in the blood. It can be done to determine the capacity of urea nitrogen in the blood.

Creatinine is a breakdown product of creatine, an important component of muscle. The creation of creation depends on muscle form, which varies very slightly. Creatinine is defecated exclusively by the kidneys, and the level in the blood is relational to the glomerular filtration rate. The serum creatinine level provides a more subtle test of kidney function than BUN because kidney impairment is nearly the only cause of high creatinine. It can also be measured with a urine test.

Creatinine clearance rate controls how professionally the kidneys are clearing creatinine from the blood and helps as an estimate of kidney function. For renal function tests, urine and serum stages of serotonin are measured, as well as the volume of urine defecated over a 24-hour period. The creatinine clearance rate is then designed and expressed as the capacity of blood, in milliliters, that can be clean of creatinine in 1 minute. A low creatinine allowance value indicates abnormal kidney function. It needs both a urine and blood sample.

2.2Fine-needleaspiration biopsy (**FNAB**, **FNA** or **NAB**), or fine-**needle aspiration cytology** (**FNAC**), is a diagnostic procedure used to investigate superficial (just under the skin) lumps or masses. In this FNAC method, a high, hollow needle is inserted into the mass for sampling of cells that, next being stained, will be inspected under a microscope. There could be a cytology test of aspiratehistologically. Fine-needle aspiration biopsies are actually safe, minor medical procedures. Often, a major surgical has beenexceptional or open biopsy can be avoided by performing a needle aspiration biopsy in its place. In the year of 1981, the first fine-needle ambition biopsy in the United States was done at Maimonides Medical Center, removing the need for surgery and hospitalization. Today, this procedure is extensively used in the diagnosis of cancer and provocative conditions.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

2.3 Biopsy - A biopsy is a test that's performed to inspect tissue or cells from a portion of the body. It can be done by scraping or cutting a small piece of the tissue or by withdrawing anexample of tissue with a needle and syringe.

2.4 Diagnostic Solography (ultrasonography) is an ultrasound-based diagnostic imaging system used for visualizing insidethe human body structures, including tendons,joints, muscle and vessels for possible pathology or lesions. The generally used, the repetition of examining pregnant women using ultrasound is called obstetric sonography.

2.5 Chest x-ray or computed tomography (CT)is Computed tomography, is known as a CT or CAT scan, is a diagnostic medicallab test that, like a traditional test of x-rays, produces multiple images or pictures of the inside of the body. The cross-sectional images generated during a CT scan can be reformatted in several planes, and can even generate three-dimensional images.These dimensional images can be viewed on a computer system monitor, printed on film or transferred to a DVD.

CT images of internal human organs, bones, soft tissue and blood vessels typically provide better detail than traditional x-rays, particularly of soft tissues and blood vessels. Using a range of techniques, including adjusting the radiation dose based on patient size and modern software technology, the quantity of radiation required to perform a chest CT scan can be significantly reduced. A low-dose chest CT produces pictures of adequate image quality to sense many lung diseases and abnormalities using up to 65 percent less ionizing radiation than a conservative chest CT scans. Low dose chest CT is presently used clinically, especially for sensing lung cancer and lung nodules. Other diseases, such as the discovery of pulmonary embolism and interstitial lung disease may not be appropriate for low-dose chest CT.

There is continuing research to further minor radiation doses. The radiologist will decide the proper settings to be used for a scan depending on a medical problem and what information is needed from the CT scan. If the child is to have a CT scan, the suitable low-dose pediatric should be applied.

III. ASSOCIATION RULE MINING

Data mining technique, the association rule mining is functional to search the hidden relations among the attributes. It classifies strong rules discovered in databases using different events of interest.Therefore, an association rule is a design that conditions when X happens, Y occurs with certain probability.

3.1 Apriori algorithm

The apriori is a classic algorithm for frequent item set mining and association rule learning over the transaction databases [9]. It continues by classifying the frequent separate items in the database and distributing them to bigger and larger item sets as long as those item sets appear sufficiently often in the database. The recurrent item sets determined by a apriori can be used to determine association rules, which hyphenate general trends in the database [10]. Association rule mining using the apriori algorithmuses a “bottom up” method, breadth-first search and a hash tree structure to count the candidate item sets efficiently. The algorithm is referred below:

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.

To select stimulating rules from the set of all possible rules created, constraints on various measures of meaning and interest can be used. The best-known constraints are lowest thresholds on provision and confidence.

3.1.1 Pseudocode for the Apriori algorithm is as follows:

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
return Ck Lk;
Where, Ck: Candidate itemset of size k
Lk: frequent itemset of size k

3.2 Rule Measures

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

3.2.1. Support: Support is defined on itemsets and gives the proportion of transactions which contain (X ∪Y). It is applied as a standard of significance (importance) of an itemset.

Since it essentially uses the count of transactions, he is often called an incidenceconstraint. An itemset with a support> a set minimum support threshold is called a frequent *or large* item set.

Supports main feature is that it possesses the *down-ward closure property (anti-monotonicity)* which means that all subsets of a frequent sight (support> minimumsupport threshold) are also frequent.

This property (actually, the fact that no superset of an infrequent set can be frequent) is used to prune the search space (usually thought of as a lattice or tree of item sets with increasing size) in the Apriori algorithm.
Supp(XUY) = P(X ∪Y).

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

supp(X)= no. of transactions which contain the itemset X / total no. of transactions

3.2.2. Confidence: Confidence is defined as the probability of seeing the rule's consequences under the condition that the transactions also contain the antecedent. Confidence is focused and gives different standards for the rules X-> Y and Y -> X.

Confidence is not down-ward closed and was established together with support by Agrawal et al. Support is first used to discovery frequent (significant) itemsets exploiting its down-ward closure property to prune the search space. Then confidence is used in a another step to produce rules from the frequent itemsets that exceed a minimum confidence threshold.
conf(X -> Y) = supp(X -> Y)/supp(X)
= P (X and Y) /P (X) = P (Y | X)

3.2.3. Lift: Lift measures how many times more often **X** and **Y** occur together than expected if they were statistically independent.

Lift is not down-ward closed and does not suffer from the rare item problematic. Also lift is susceptible to noise in minor databases. Rare itemsets with low counts (low probability) which perchanceoccurs a few times (or only once) together can produce enormous lift values.

lift(X -> Y) = lift(Y -> X) = conf(X -> Y)/supp(Y) = conf(Y -> X)/supp(X) = P(X and Y)/(P(X)P(Y))

3.2.4. Leverage: Leverage measures the difference of **X** and **Y** appearing together in the data set and what would be expected if **X** and **Y** were statistically dependent. The balanced in a sales setting is to find out how many more units (items **X** and **Y** together) are sold than predictable from the independent cells.
leverage(X -> Y) = P(X and Y) - (P(X)P(Y))

3.2.5. Conviction: Conviction was developed as an alternative to confidence which was found to not capture the direction of relations adequately. Conviction compares the probability that **X** appears without **Y** if they were dependent with the actual frequency of the appearance of **X** without **Y**. In that respect it is similar to lift, however, it contrast to lift it is a directed measure since it also uses the information of the absence of the consequent. An exciting fact is that the principle is monotone in confidence and lift.

Conviction (X -> Y) = (1-Supp (Y)) /(1-conf (X -> Y)) = P (X) P (not Y) /P (X and not Y)

IV. CONCLUSION

The association rule mining algorithm, apriori has recognized the position of investigations and laboratory tests in measuring the degree and extent of oral cancer. The rules clearly show that if FNAC of neck node, USG and CT scan or MRI is positive then the chance of survival is concentrated. However, if LFT is normal, the probability of survival is high. If a site is either tongue, plate or BM and diagnostic biopsy are Squamous–Cell-Carcinoma then analysis usually is SCC then it obviously indicate oral cancer, which may lead to high humanity if appropriate action is not introduced. In future, intend to extend this article by trying to extract important patterns and useful rules through the association rule mining algorithm from extracting most effective course of action.

REFERENCES

[1] Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996) Data Mining to Knowledge Discovery in Databases. AIMagazine, **17**, 37-54.
[2] Han, J., Kamber, M. and Pei, J. (2011) Data Mining: Concepts and Techniques. 3rd Edison, Morgan Kaufmann Publishers,Burlington.
[3] Khosla, R. and Dillon, T. (1997) Knowledge Discovery, Data Mining and Hybrid Systems. Engineering, IntelligentHybrid Multi-Agent Systems, Kluwer Academic Publishers, Norwell, 143-177.
[4] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996) Data Mining to Knowledge Discovery: An Overview. Advancesin Knowledge Discovery and Data Mining, AAAI Press/MIT Press, 1-36.
[5] Cios, K.J. (2001) Medical Data Mining and Knowledge Discovery. Studies in Fuzziness and Soft Computing, **60**, 502.
[6] K.R.Coelho, "Challenges in Oral Cancer Burden in India," Journal of Cancer Epidemiology, vol. 2012, Article ID 701932, 17 pages.
[7] Elango, J.K., Gangadharan, P., Sumithra, S. and Kuriakose, M.A. (2006) Trends of Head and Neck Cancers in Urban And Rural India. Asian Pacific Journal of Cancer Prevention, **7**, 108-112.
[8] American Cancer Society (2012) Cancer Facts and figures, Atlanta (GA), The Society.
[9] Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago de Chile, 12-15 September 1994, 487-499.
[10] Zaki, M.J. (2013) Mining Non-Redundant Association Rules. Data Mining and Knowledge Discovery, **9**, 223- 248. http://dx.doi.org/10.1023/ B:DAMI.0000040429.96086.c7