ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

Lung Cancer Early Diagnosis Using Some Data Mining Classification Techniques: A Survey

Thangaraju P[1], Barkavi G[2]

[1]Asst. Professor, [2]M.Phil, Scholar
Department of Computer Applications,
Bishop Heber College (Autonomous),
Trichirappalli-620 017

Abstract: Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is primarily used to this requirement thus finding its applications in diverse fields such as retail, financial, communication, marketing organizations and medicine. Data Mining plays an important role in healthcare organization because with the growth of population and dangerous deadly diseases like Cancer, SARS, Leprosy, HIV etc, Lung cancer is one of the most dangerous disease. This survey for appropriate medical image mining, Data Preprocessing, Feature Extraction, rule generation and classification, it provides basic framework for further improvement in medical diagnosis.

Keywords: *Data mining, lung cancer, early diagnosis, classification, neural networks.*

1. Introduction

Uncontrolled cell growth in tissues of the lung is called Lung cancer disease, There are two types of lung cancer. Non small cell lung cancer (NSCLC) and small cell, lung cancer (SCLC) or oat cell cancer [2]. If it is not diagnosed and treated early, the tissues can be metastasized to other parts, early stage Detection of lung cancer is the key of its cure there is no tool for detecting cancer in early stage is accepted by cancer care. lung cancer diagnosis using current tools like X-ray chest films, CT, MRI[1] and author not only high cost it efficient in only stage four and it cause pain to the patients also CT scan performed 0.4 % of current scans as per 2007 report[3].

Biomarker lung cancer can also help in lung cancer but there is no specify biomarkers, researchers are working on. Medical imaging an essential way for diagnosis and treatment these are different from typical photographic images, these images includes both projection X-Ray chest film and cross-sectional images [3].

Medical data mining is a promising place for applied to an analyze patients' records automatically aiming at the discovery of new knowledge for decision making, Knowledge Discovery in Databases (KDD). In general, Data Mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as Data Mining is one of the most important stages of the KDD process [4]. This knowledge is used to disease diagnosis and successful treatment. This paper methods classify digital X-ray chest films into two categories they are abnormal and normal. The normal ones are those characterizing a healthy patient. Abnormal ones include type of lung cancer, using some data mining, techniques such as neural networks and association rule mining.

2. Data Mining Techniques

2.1 Data Preprocessing

To improve the quality of images its necessary to Preprocessing phase of the images and feature extraction phase make more reliable. This phase made up of some processes they are data normalization, data preparation, data transformation, data cleaning, and data formatting. Data normalization process is necessary to combine the different image formats to a regular format.

Data preparation modifies the image to suitable format for transformation techniques. The image will transformed in order to obtain a compressed representation. Segmentation completed to recognize regions of interest (ROI) for the mining task usually achieved using classifier systems. This step finds consequent regions within an image, since item sets are extremely large [3].

2.2 Feature Extraction

Images have large number of features, To decrease the complexity of processing it is important to recognize and extract interesting features for an exacting task in order. All the images of attributes are not useful in knowledge extraction. Image extract attribute is used for automatically extract image attributes like local color, global color, texture, and structure [1]. This extraction of images required for many image mining applications such as content based information retrieval (CBIR), image classification etc. It localizes the extraction process to very small regions in order to ensure that capture all areas[5].

2.3 Rule Generation

Medical images are not self contained so data integration is an important and often used in a combination with other patient data in the process of diagnosis. suppose that association rules of two forms they are (1) Image contents dissimilar to spatial relationships, e.g., if an image has a texture X, that it is likely containing protrusion Y; (2) Image contents associated to spatial relationships, e.g., if X is among Y and Z it is likely there then there is a T beneath. A low minimum support and high minimum confidence is pleasing, since few image datasets have high support value.

2.4 Data Set

In this paper, assume training dataset as the 300 x-ray chest films multimedia database used in proposed classification system. The mentioned database contains a real data values in from of x-ray chest images. Also consider 70 percent as a training value of the systems and 10percent for testing it. Ten splits of the data collection will considered to compute all the results in order to obtain a more accurate result of the system potential [1].

2.5 Classification

Data mining retrieve data's models by investigative already classified data and inductively finding a relating pattern. Now a days many advanced classification approaches, such as neural networks, fuzzy-sets, and expert systems, have been widely applied for image classification. Image classification grouped as parametric and nonparametric, or hard and soft (fuzzy) classification, or per-pixel, sub pixel, and per field [1]. The most regularly used non-parametric classification approaches are neural networks, decision trees, SVM, and expert systems [8].

2.6 Neural Networks

ANN has proven to be a useful tool in pattern recognition and classification tasks in diverse areas, including clinical medicine [10]. Despite the wide applicability of ANN, the large amount of data required for training makes using them an unsuitable classification technique when the available data are scarce. The scarcity of data and the complexity of interpretation of relevant physiological information impose extra demands that prohibit the applicability of most statistical and machine learning techniques developed [11]. The structure of the neural network consists of three layers: the input layer, the hidden layer and the output layer. The number of nodes in the input layer is equal to the number of elements existing in one transaction in the database [11].The nodes in the hidden layer may Connect to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables [9].An NN is adaptive in nature because it changes its structure and adjusts its weight in order to minimize the error. Adjustment of weight is based on the information that flows internally and externally through network during learning phase. This concept drives us to modify the interior weights while trained neural network used to classify new images.

Steps Performed in Neural Network Classifier:

- Create feed-forward back propagation network.
- Train neural network with the training samples and the group defined for it.
- The input image extracted PCA standardized data as the test samples, simulate the neural network to check whether the particular selected input sample has cancer or not.
- From the results of network and the samples trained in network classification rate is calculated using some mathematical formulas [9].

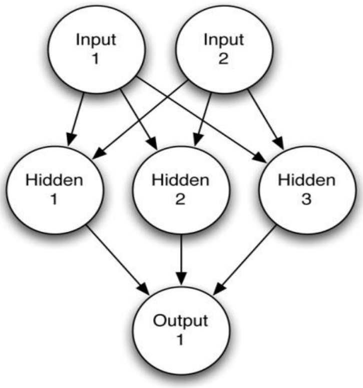

Fig.1 A neural network with one hidden layer

2.7 Association Rule Mining

Association rule mining has been commonly investigated in data mining literatures. Many proficient algorithms have been proposed, one of the most popular algorithms are called Apriori and FP Tree growth Association rule mining classically intends at discovering associations between items in a transactional database." Given a set of transactions D = {T1,…… ,Tn} and a set of items I = {I1,I2…….,im} such that any transaction T in D is a set of items in I, an association rule is an implication A → B where the antecedent A and the consequent B are subsets of a transaction T in D, and A and B have no common items. For the association rule to be acceptable, the conditional probability of B given A has to be higher than a threshold called minimum confidence "[7].

2.8 Appendix: Apriori algorithm

Several algorithms have been developed (Chen *et al* 1996) aiming to extract ARs. The *Apriori* algorithm (Agrawal and Srikant 1994) is a state-of-the-art algorithm since most of the AR algorithms are variations of this. It works iteratively finding first the set of large 1-itemsets, and then set of 2-itemsets, and so on. The number of scans depends on the length of the maximal item set. *Apriori* is based on the generation of a smaller candidate set using the set of large item sets found in the previous iteration. Let Lk represent the set of frequent k-item sets and Ck the set of candidate itemsets or potentially frequent item sets. The *Apriori* algorithm makes many passes over the SPECT database and each pass consists of two stages. In the first one, the set of all frequent ($k-1$)-itemsets, $Lk-1$, found in the ($k-1$)st pass, is used to generate the candidate item sets Ck which are ensured to be a superset of the set of all frequent k-item sets.

Secondly, the algorithm scans the database and it decides which of the candidates in Ck are contained in the record using a hash-tree data structure and incrementing their support count. Finally of the pass, Ck is tested to determine which of the candidates are frequent, yielding Lk. The *Apriori* algorithm finishes when Lk becomes empty (Skirant *et al* 1997)[6].

We will use Apriori algorithm to discover association rules among the features extracted from the x-ray chest films database and the category to which each x-ray chest films belongs.

3. Conclusion

In this paper we are using some data mining classification techniques like neural network and association rule mining for detection and classification Lung Cancer in X-Ray chest films. 300 x-ray chest films multimedia database as a training dataset used in our proposed classification system. We Classify the digital X-ray chest films in two categories: normal and abnormal. The normal ones are those characterizing a healthy patient. The abnormal ones include Types of lung cancer. We will use some procedures as a essential part to the task of medical image mining theses procedures includes Data Preprocessing, Feature Extraction and Rule Generation. In this paper we well use classification methods in order to classify problems aim to identify the characteristics that indicate the group to which each case belongs.

4. References

[1] Zakaria Suliman Zubi and Rema Asheibani Saad, “Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer,” Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases, Libya, 2007.

[2] V.Krishnaiah, Dr.G.Narsimha and Dr. N.Subhash Chandra,” Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques,” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, pp. 39 – 45.

[3] Juliet R Rajan and Jefrin J Prakash, “Early Diagnosis of Lung Cancer using a Mining Tool,” National Conference on Architecture, Software systems and Green computing-2013(NCASG2013).

[4] Divya Tomar and Sonali Agarwal, “A survey on Data Mining approaches for Healthcare,” International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266

[5] Jaba Sheela L and Dr.V.Shanthi “An Approach for Discretization and Feature Selection Of Continuous-Valued Attributes in Medical Images for Classification Learning,” International Journal of Computer Theory and Engineering, Vol. 1,No.2, June2009 1793-8201, pp. 154.

[6] R Chaves, J M G´orriz, J Ram´ırez, I A Ill´an, D Salas-Gonzalez and M G´omez-R´ıo “Efficient mining of association rules for the early diagnosis of Alzheimer’s disease,” Institute of Physics and Engineering in Medicine 2011.

[7] Zakaria Suliman Zubi, Rema Asheibani Saad, “Improves Treatment Programs of Lung Cancer Using Data Mining

Techniques,"Journal of Software Engineering and Applications, 2014, 7, pp. 69-77.

[8] Ada and Rajneet Kaur, " A Study of Detection of Lung Cancer Using Data Mining Classification Techniques," International Journal of Advanced Research in Computer Science and Software Engineering march 2013.

[9] Ada and Rajneet Kaur, "Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier." International Journal of Application or Innovation in Engineering & Management June 2013.

[10] Lakhmi Jain and Philippe De Wilde," Practical Applications of Computational Intelligence Techniques" ISBN 0-7923-7320-0, pp. 300.

[11] Turban "Decision Support And Business Intelligence Systems, 8/E," Pearson Education India, 01-Sep-2008 ISBN 978-81-317-2425-5, pp. 373.