Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer

ZAKARIA SULIMAN ZUBI[1], REMA ASHEIBANI SAAD[2]

[1]SIRTE UNIVERSITY, FACULTY OF SCIENCE, COMPUTER SCIENCE DEPARTMENT
SIRTE, P.O BOX 727, LIBYA,
{zszubi@yahoo.com}
[2]ALFATEH UNIVERSITY, FACULTY OF SCIENCE, COMPUTER SCIENCE DEPARTMENT, TRIPOLI, LIBYA,
P,O BOX 13210,
{ranem_reem@yahoo.com}

Abstract: - Lung cancer is a disease of uncontrolled cell growth in tissues of the lung, Lung cancer is one of the most common and deadly diseases in the world. Detection of lung cancer in its early stage is the key of its cure. In general, a measure for early stage lung cancer diagnosis mainly includes those utilizing X-ray chest films, CT, MRI, etc. Medical images mining is a promising area of computational intelligence applied to automatically analyzing patient's records aiming at the discovery of new knowledge potentially useful for medical decision making. Firstly we will use some processes are essential to the task of medical image mining, Data Preprocessing, Feature Extraction and Rule Generation. The methods used in this paper work states, to classify the digital X-ray chest films into two categories: normal and abnormal. The normal state is the one that characterize a healthy patient. The abnormal state including the types of lung cancer; will be used as a common classification method indicating a machine learning method known as neural networks. In addition, we will investigate the use of association rules in the problem of x-ray chest films categorization.

The digital x-ray chest films are storied in huge multimedia databases for a medical purpose. This multimedia database provides a great environment to apply some *image recognition methods* to extract the useful knowledge and then rules from the mentioned database. These rules that we could got using image recognition methods, will help the doctors to decide important decisions on a particular patient state.

Key-Words: - Data Mining, Classification, Medical Imaging, Image Recognition, Neural Networks, Association Rule Mining, Early Cancer Diagnosing.

1 Introduction

According to the American Lung Association (2005), lung cancer is presently the leading cause of death from cancer in the United States [4]. Lung cancers classified into two main categories: small-cell lung cancers (SCLC), which report for about 20% of cases, and non-small-cell lung cancers (NSCLC), which report for the other 80%. Non-small-cell lung cancers include squamous cell carcinomas (35% of all lung cancers), adenocarcinomas (27%) and large cell carcinomas (10%) [3].

Currently, there are no technical methods to prevent lung cancer, which is why early detection represents a very important factor in cancer treatment and allows attainment a high survival rate. Medical images as an essential part of medical diagnosis and treatment were concentrating on these images for goods. These images are different from typical photographic images primarily because they make known internal anatomy as opposed to an image of surfaces.

These images include both projection X-Ray chest film and cross-sectional images. Those types of images acquired by means of computed Tomography (CT) or magnetic resonance imaging (MRI), or one of the other tomography modalities (SPECT, PET, or ultrasound, for example). Medical image processing is a Branch of image processing that deals with such images. It is driven both by the peculiar nature of the images and by the medical applications that make them useful. Medical images include prosperity of unseen information that exploited by physicians in making reasoned decisions about a patient. However, extracting this

relevant hidden information is a critical first step to their use. This reason motivates us to use data mining techniques capabilities for efficient knowledge extraction [2].

As a first glance, Lung X-ray chest films considered as the most reliable method in early detection of lung cancers, while the accuracy rate in a high number of x-ray chest films read by physicians tends to be decreases day by day. The automatic reading of digital x-ray chest films turns out to be highly popular. It has been verified that dual reading of lung x-ray chest films (consecutive reading by two physicians or radiologists) improved the accuracy rate, but at high costs. According to the fact that the medical domain involves high accuracy and particularly the rate of false negatives are very low. In addition, another significant factor that influences the success of classification methods is working in a team with medical specialists, which is attractive but often not achievable [8]. That is why the computer aided diagnosis systems are necessary to support the medical staff to achieve high capability and effectiveness. This is the main reason for the development of classification systems to Diagnosing Lung Cancer.

Image recognition mining deals with the extraction of image patterns from a large collection of images stored in particular multimedia databases. Obviously, image mining is different from low-level computer vision and image processing techniques because the focus of image mining is in extraction of patterns from large collection of images, whereas the focus of computer vision and image processing techniques is in understanding and/or extracting specific features from a single image. Although there looks like to be some overlaps between image mining and content-based retrieval (both are dealing with large collection of images), image mining goes beyond the problem of retrieving relevant images. In image mining, the aim is the discovery of image patterns that are considerable in a given collection of images [6].

Mining medical images involves many processes. Medical data mining is a promising area of computational intelligence applied to an automatically analyze patients' records aiming at the discovery of new knowledge potentially useful for medical decision-making. Induced knowledge is anticipated not only to increase accurate diagnosis and successful disease treatment, but also to enhance safety by reducing medication-related errors.

The methods in this paper 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 a common classification method, namely neural networks, but significantly improve the accuracy rate of the classifier compared to other published results using the same dataset. In addition, we investigate the use of association rules in the problem of x-ray image categorization and demonstrate with encouraging results that association rule mining is a promising alternative in medical image classification and certainly deserves more attention as well.

2 The objectives of the paper

The aims of this paper work are pointed out as follow:

1. Using some data mining, techniques such as neural networks and association rule mining techniques to detection early Lung Cancer and to classified it by using X-Ray chest films image.

2. We will use the data of 300 x-ray chest films as a dataset in our classification system.

3. 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.

4. Helping physicians to decide an important decisions on a particular patient state.

3 Data Mining Task

We will use some essential processes as necessary requirements to assign medical image mining mission:

Data Preprocessing:

Preprocessing phase of the images is necessary to improve the quality of the images and make the feature extraction phase more reliable. This phase consists of some processes. These processes contain data normalization, data preparation, data transformation, data cleaning, and data formatting. Normalization techniques are necessary to combine

ISBN: 978-960-474-273-8

the different image formats to a regular format. Data preparation modifies images to present them in an appropriate format for transformation techniques. The image will transformed in order to obtain a compressed (lossless) representation of it, e.g., using wavelet transforms. Segmentation completed to recognize regions of interest (ROI) for the mining task usually achieved using classifier systems. The segmentation step finds consequent regions within an image, since item sets are extremely large [5].

Feature Extraction:

Images usually have a huge number of features. It is important to recognize and extract interesting features for an exacting task in order to decrease the complexity of processing. These are attributes or portion of the image being analyzed that is probably to give interesting rules for that problem. Not all the attributes of an image are useful for knowledge extraction. This stage raises the overall effectiveness of the system. Image processing algorithms used, which automatically extract image attributes such as local color, global color, texture, and structure. Texture is the mainly useful description property of an image and it specifies attributes, such as resolution, which used in image mining [5].

Feature extraction from images are required for many image mining applications such as content based information retrieval (CBIR), image classification etc. These features typically extracted based on the image's information by image processing only [10]. An image can satisfactorily characterize using the attributes of its features. The extraction of the features from an image can finished using a variety of image processing techniques. We localize the extraction process to very small regions in order to ensure that we capture all areas.

Rule Generation:

In an extremely knowledge based domain associated with a domain knowledge we can progress some data-mining tasks to improve decision support services [5]. This data integration is an important notion because medical images are not self contained, and often used in a combination with other patient data in the process of diagnosis. We suppose that association rules of two forms such as: (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 [5].

4 Data Set

In our paper, we will consider the 300 x-ray chest films multimedia database as a training dataset used in our proposed classification system. The mentioned database contains a real data values in from of x-ray chest images. We will also consider 70 percent as a training value of the systems and 10 percent 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.

5 Methods and Models

There are several kinds of data mining methods; some of the major data mining methods are known as deviation detection, summarization, classification, generalized rule induction and clustering [1].

In this paper work, we well use classification and generalized, neural network and association rule mining induction methods in order to classify problems aim to identify the characteristics that indicate the group to which each case belongs.

Classification and Generalized

Data mining generates classification models by investigative already classified data (cases) and inductively finding a predictive pattern. These accessible cases may come from a chronological database, such as people who have already undergone a meticulous medical treatment or moved to a new long distance service. They may come from a conduct experiment in which a sample of the whole database tested in the real world and the results used to produce a classifier. For example, a sample of a mailing list would send as an offer, and the results of the mailing used to build up a classification model to apply to the entire database. Sometimes an expert classifies a sample of the database, and this classification is then used to make the model, which will be useful to the whole database [2]. In recent years, many advanced classification approaches, such as neural networks, fuzzy-sets, and expert systems, have been widely applied for image classification. In most cases,

ISBN: 978-960-474-273-8

image classification approaches grouped as supervised and unsupervised machine learning approaches, or parametric and nonparametric, or hard and soft (fuzzy) classification, or per-pixel, subpixel, and per field [4].

Per-pixel classification algorithms can be parametric or non-parametric. The parametric classifiers assume that a normally distributed dataset exists, and that the statistical parameters (e.g. mean vector and covariance matrix) produced from the training samples are representative. However, the hypothesis of normal spectral distribution often dishonored, especially in complex landscapes. In addition, insufficient, non-representative, or multimode distributed training samples can further establish uncertainty to the image classification procedure. Another major negative aspect of the parametric classifiers lies in the difficulty of integrating spectral data with auxiliary data.

The maximum probability may be the most regularly used parametric classifier in practice, because of its robustness and its easy accessibility in almost any image-processing software. With non-parametric classifiers, the assumption of a normal distribution of the dataset is not required. No statistical parameters are required to separate image classes. Non-parametric classifiers are accordingly especially appropriate for the incorporation of non-spectral data into a classification procedure. The most regularly used non-parametric classification approaches are neural networks, decision trees, support vector machines, and expert systems [4].

Neural Network

Especially, the neural network approach has been widely adopted in recent years. The neural network has several advantages, including its nonparametric nature, arbitrary decision boundary capability, easy adaptation to different types of data and input structures, fuzzy output values, and generalization for use with multiple images. Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions.(Actual biological neural networks are incomparably more complex.) Neural nets may used in classification problems (where the output is a categorical variable) or for regressions (where the output variable is continuous).

The architecture of the neural network shown in figure1 consists of three layers such as input layer, hidden layer and output layer. The nodes in the input layer linked with a number of nodes in the hidden layer. Each input node joined to each node in the hidden layer. 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 [2].

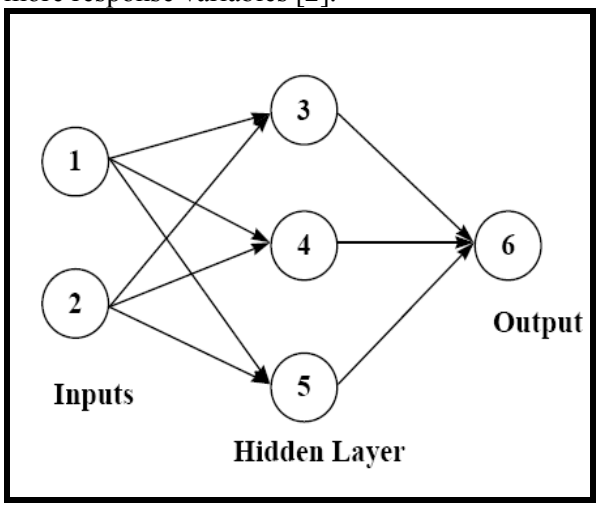

Figure 1. A neural network with one hidden layer.

A main concern of the training phase is to focus on the interior weights of the neural network, which adjusted according to the transactions used in the learning process. For each training transaction, the neural network receives in addition the expected output [8]. This concept drives us to modify the interior weights while trained neural network used to classify new images.

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 $D = \{T_1,\ldots\ldots ,T_n\}$ and a set of items $I = \{I_1,I_2\ldots\ldots, i_m\}$ such that any transaction T in D is a set of items in I, an association rule is an implication $A \rightarrow B$ 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 "[2].

Association rules mining is usually a two-step process, wherein the first step frequent item-sets are

discovered (i.e. item-sets whose support is no less than a minimum support) and in the second step, association rules derived from the frequent item sets. Most algorithms used to identify large itemsets can classify as either sequential or parallel.

Ordinarily, it understood that the itemsets identified and stored in lexicographic order (based on item name). This ordering provides a logical manner in which itemsets can be generate and counted. This is the normal approach with sequential algorithms. On the other hand, parallel algorithms focus on how to parallelize the task of finding large itemsets [6].

The Apriori algorithm called also as "***Sequential Algorithm***" developed by [Agrawal1994] is a great accomplishment in the history of mining association rules [Cheung1996c]. It is also the most well known association rule algorithm. This technique uses the property that any subset of a large itemset must be a large item set it self. In addition, it assumed that items within an itemset reserved in lexicographic order. The primary differences of this algorithm from the AIS and SETM algorithms are the way of producing candidate itemsets and the selection of candidate itemsets for counting. As stated earlier, in both the AIS and SETM algorithms, the ordinary itemsets between large itemsets of the previous pass and items of a transaction achieved. These ordinary itemsets extended with other individual items in the transaction to generate candidate itemsets. However, those individual items may not be large. As we know that a superset of one large itemset and a small itemset will result in a small itemset, these techniques produce too many candidate itemsets, which turn out to be small. The Apriori algorithm addresses this important issue and generates the candidate itemsets by joining the large itemsets of the previous pass and deleting those subsets, which are small in the previous pass without considering the transactions in the database. By only making an allowance for large itemsets of the previous pass, the number of candidate large itemsets is considerably reduced [7].

In our approach, we will use the apriori algorithm in order 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.

6 Paper Outline

In the first section of the paper, we have introduced lung cancer. The basic ways to diagnose this disease using medical images and special X-ray images, and how are the difficulties and expensive the reading of this topic images?, This led us to think about the possibility of introducing computer algorithms solution using Data Mining Techniques. The techniques help us in the development of early diagnostic system to support Doctors in making their decisions in the diagnosis of the disease, which leads to a quick recovery.

In the second section, we set out objectives aim of this paper, which we will look forward to achieve it.

In the third section, we have identified the basic processes, which we will undertake to perform the functions of data mining. This includes data preprocessing phase that contain a combination of operations, including data transformation and normalization of the data.

The section derives us to the stage of Rule Generation, where the images have a large number of essential features that must extract to represent important features of our work. It releases the complexity of the treatment process as an important benefit in this work. Next stage, followed by generation rules, as the data integrity is very important, especially in the medical images, we have assumed association rules in two forms.

In the fourth section, we have recognized data set that we will use in this paper.

In the fifth section, we indicates the most essential methods used in data mining, and focused on the classification method that we are willing to use in this paper, techniques of this method and how they work are mentioned as well.

7 Conclusion

We will summarize the main concern of this paper work in the following manner:

1. In this paper, we are going to use some data mining, techniques such as neural networks and association rule mining techniques, for detection and classification Lung Cancer in X-Ray chest films.

2. We will consider the 300 x-ray chest films multimedia database as a training dataset used in

our proposed classification system.

3. From these set of images we will consider 70 percent as a training value of the systems and 10 percent 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.

4. 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.

5. 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.

6. In this paper work, we well use classification and generalized, neural network and association rule mining induction methods in order to classify problems aim to identify the characteristics that indicate the group to which each case belongs.

References:

[1]. Alex A. Freitas, A Genetic Programming Framework for Two Data Mining Tasks: Classification and Generalized Rule Induction, Page 1.

[2] by Two Crows Corporation Introduction to Data Mining and Knowledge Discovery .Third Edition,2005. ISBN: 1-892095-02-5, Pages 10, 11.

[3].Developed by the National Collaborating Centre for Acute Care,Lung cancer
The diagnosis and treatment of lung cancer,ISBN: 1-84257-920-7 ,Published by the National Institute for Clinical Excellence February 2005. Page 4.

[4] D. LU* and Q. WENG. A survey of image classification methods and techniques for improving Classification performance. International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online # 2007 Taylor & Francis, Page 829

[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, Page 154.

[6] Ji Zhang Wynne Hsu Mong Li Lee Image Mining: Issues, Frameworks and Techniques,Page1.

[7]Margaret H. Dunham, Yongqiao Xiao Le Gruenwald, Zahid Hossain A SURVEY OF ASSOCIATION RULES. Page 9.

[8] Maria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coman Application of Data Mining Techniques for Medical Image Classification. Page 97.

[9] Qingyu Zhang, Richard S. Segall VISUAL ANALYTICS OF MINING HUMAN LUNG CANCER DATA, Page 1.

[10] Tianxia Gong, Chew Lim Tan, Tze Yun Leong, Cheng Kiang Lee, Boon Chuan Pang, C. C. Tchoyoson Lim, Qi Tian, Suisheng Tang, Zhuo Zhang, Text Mining in Radiology Reports. page 4.