UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Faculty of Computer Science and Information Technology

Ensemble Method for Classification Mammogram Images

By

Alaa abayzeed Mohmmed

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements

in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Technology

Supervisor

Dr. Ali Ahmed Alfaki Abdallah

September 2016

Abstract

Brest Cancer is the disease of the age and it threat women's in the developed country by finding early detect technique that help the doctors and specialist to reduce the number of death by cancer. Classification is the organization of data in given classes, the classification uses given class to order the objects. The most important part in classification approaches or model is the classification algorithm which used to learn from the training set and build the model. The main reason of low accuracy of most previous studies comes as a result of using either not accurate features or not proper classification method, this study built a ensemble classifier by uses six features that extracted from MIAS dataset, the dataset splits into two parts training and testing, after learn the classifier based on training set then the classifier tested based on the test set to determine the accuracy of the classifier depend on the confusion matrix. This study emphasis of five phases starting in collect images, preprocessing, features extracting, classification and end with testing and evaluating. The result of the proposed method comes as 76% accuracy in training and testing data . We recommend to use more than six features and apply it into the decision tree model.

المستخلص

يعتبر مرض سرطان الثدى من اخطر امراض العصر التى تهدد النساء خصوصا في الدول النامية ،والى الان لم يتم تحديد واكتشاف اسباب المرض،ايجاد الالية للاكتشاف المبكر ستساعد الاطباء والباحثين في التقليل من عدد الوفيات من هذا المرض.تهدف هذه الدراسة لبناء الية تساعد الاطباء علي سرعة وسهولة التصنيف بدقة. عملية تصنيف وترتيب البيانات تعتبر علية تنظيم وترتيب للبيانات،واهم جزئية فى عملية التصنيف وبناء نموذج التصنيف هى اختيار خوارزمية التصنيف المناسبة والتى تستخدم لتعليم المصنف السلوك المطلوب.اهم عوامل انخفاض دقة المصنفات ناتجة عن استخدم مصنف غير مناسب اواستخلاص خصائص غير مناسبة مع البيانات او غير كافية.تستخدم هذه الدراسة مصنف المجموعات كاداة لتصنيف البيانات وتم استخدام ستة خصائص تم استخلاصها من صور اشعة لصدر الانسان ماخوذة من جمعية تحليل صور سرطان الثدى. احتوت هذه الدراسة علي خمسة مراحل بداية بمرحلة جمع الصور ثم تهيئتها واستخلاص الخصاتص ومن ثم مرحلة التصنيف انتهاءا بمرحلة الاختبار والتقييم.قى مرحلة الاختبار اظهر نموذج التصنيف نتائج وصلت ل 76 %.

يوصى باستخدام خصائص اخري غير المستخدمة فى هذه الدراسة لبناء نموذج التصنيف فى الدراسات المستقبلية.

1.introduction

The recent revolution in computer technologies and storage volume have produced huge amount of data and information from many sources such as social networks, online databases, engineering systems, and health information systems. At the present time, many countries around the world are changing the way to apply health care to the patients and the people by utilizing the benefits of advancements in computer technologies through Medical images technologies used digital mammography.

1.1 Concept of Data mining

Data mining is the process of discovering interesting patterns from massive amounts of data. As a knowledge discovery process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation.[1]

Data Mining is also known as Knowledge Discovery, Knowledge Extraction, Data/Patterns Analysis, Data Archeology, Information Harvesting and Business Intelligence. Data mining can often be successfully applied to these problems by improving the efficiency of the systems and the designs of the machines. Numerous data mining applications, involve tasks that can be set up as supervised.[2]

Every day the human beings are using the vast data and these data are in the different fields .It may be in the form of documents, may be graphical formats ,may be the video ,may be records (varying array) .As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data .As and when the customer will required the data should be retrieved from the database and make the better decision .This technique is actually we called as a data mining or Knowledge Hub or simply KDD(Knowledge Discovery Process).The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data[2]

Data mining applications in health can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored,

prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications.[2]

1.2 Data Preprocessing

Data have quality if they satisfy the requirements of the intended use. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability, Low-quality data will lead to low-quality mining results.

1.3 Classification

Classification is a form of data analysis that extracts models describing important data classes. Classification in data mining technique is used to predict group membership for data instances. Data mining involves the use of sophisticated data analysis tools to discover the relationships in large data set. Data mining never means a collection and managing data, it also includes analysis and prediction of data. Classification techniques are capable of processing a wider variety of data than regression and also growing in popularity.

People are often prone to making mistakes during analysis while establish relationships between multiple features. This makes it difficult for them to find solutions for certain problems. Data mining can often be successfully applied to these problems by improving the efficiency of the systems and the designs of the machines. Numerous data mining applications, involve tasks that can be set up as supervised. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example, we can build a classification model to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large. Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large amounts of disk-resident data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

1.4 Research Background

Automatic diagnostic systems are an important application of analysis of database and pattern recognition, aiming at assisting physicians in marking diagnostic decisions.[1] Automated diagnosis is especially used to diagnose the variety of cancers. Classification is the organization of data in given classes. The most important part in classification approaches or model is the classification algorithm which used to learn from the training set and build the model

Breast cancer classification divides into categories according to different schemes, each based on different criteria and serving a different purpose. The major categories are the histopathological type, the grade of the tumor, the stage of the tumor, and the expression of proteins and genes. As knowledge of cancer cell biology develops these classifications are updated.

The purpose of classification is to select the best treatment. The effectiveness of a specific treatment is demonstrated for a specific breast cancer (usually by randomized, controlled trials). That treatment may not be effective in a different breast cancer.

1.5 Problem Statement

The importance reason of the low accuracy of most previous studies comes as a result of using either not accurate features or not a proper classifier method. Most of the breast cancer classification method use different set of medical image features. The accuracy of classification result depends on the classifier method used. This study using six statistical features function extracted from each medical image and Ensemble classifier as a classifier.

1.6 Objectives of the Study

i. To get Region of Interest(ROI) from each mammogram images.

ii. To extract mean, standard deviation, smoothness, contract, kurtosis and sleekness features for medical image

iii. To apply ensemble classifier from based classification model from mammogram image classification.

iv. To evaluate our results compared by the previous studies.

1.7 Significant of Study

The high incidence of breast cancer in women, especially in developed countries, has increased significantly in the last years. Though much less common, the etiologies of this disease are not clear and neither are the reasons for the increased number of cases. Currently there are no methods to prevent breast cancer, which is why early detection represents as a very important factor in cancer treatment and allows reaching a high survival rate. Mammography is considered the most reliable method in early detection of breast cancer. Due to the high volume of mammograms to be read by physicians, the accuracy rate tends to decrease, and automatic reading of digital mammograms becomes highly desirable. It has been proven that double reading of mammograms (consecutive reading by two physicians or radiologists) increased the accuracy, but at high costs. That is why the computer aided diagnosis systems are necessary to assist the medical staff to achieve high efficiency and effectiveness.

1.8 Research Scope

This study covers the offline not online classification and considers the Mammogram images that taken from MIAS Data set. The evaluated measures that will used in this thesis is the Confusion Matrix (true positive, true negative, False Positive and False negative) to determine and examine the accuracy of the classifier that is used during the study.

1.9 Thesis Organization

Chapter one a general definition about data mining and its functionality also describe the problem statement of the study and objective, significant , expected result and the scope of the study. Chapter two Literature review of classification method, medical image classification. Chapter three describe the research methodology, the four phases of the study and materials that used in the study. Chapter four Describes the implementation of the ensemble classifier, build the classifier and run it in training data, test it after that in test data to determine the accuracy of the classifier. Chapter five gives the conclusion and recommendation of the study.