UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Faculty of Computer Science and Information Technology

K-Nearest Neighbor Classification Method for Mammogram Images

By

Tahani Khidir Abdelrahim

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements
In PartialFulfillment of the Requirement for the Degree of Master of Science in Information Technology

Supervisor

Dr: Ali Ahmed

September 2016

Abstract

The Breast Cancer is one of the most important causes of death in the countries of the world to women. When early detection of the disease using mammography image and features extraction of the image and then uses algorithms for the classification of these images.

Classification is data mining technology, which places the data in groups to target groups, the process of building the classification model is very important in the selection of appropriate algorithm, is divided into two sections, in the first section training data is used to build the classifier while the second section testing is used to prove the accuracy of the classification algorithm.

The classification process includes five basic steps beginning with mammogram Image collection, image processing, and the extraction of the feature of the images, classification, testing and evaluation. This study used six features, extracted from mammogram images that taken from MAIS database, then the research apply K -nearest neighbors (KNN) classification method.

This research, the dataset split into two parts, namely: training and testing. After the construction of the classifier based on training data for testing the proposed model using the test data to measure the accuracy based on the value of the (k). The best accuracy obtained 76.4% when using percentage of 85% and 15% for training and testing respectively and K value is 1. This research recommends uses other features that may add more power for the accuracy of the classifier to enhance the results.

المستخلص

يعتبر سرطان الثدي أحد أهم أسباب الوفاه لدي النساء في العالم .عند إجراء الكشف المبكر لهذا المرض يتم أستخدام صور الأشعة لتصوير الثدي وأستخلاص الخصائص من الصور ومن ثم نستخدم خوارزميات لتصنيف هذه الصور .

التصنيف هو تقنية تنقيب البيانات الذي يضع البيانات في مجموعات لأستهداف فئات محدده ، عملية بناء نموذج التصنيف مهمة جداُ في أختيار الخوارزمية المناسبة ، يتم تقسيم البيانات الي قسمين ،القسم الأول يستخدم في بناء النموذج وتعليمه أما القسم الثاني أختبار النموذج لإثبات دقة التصنيف .

عملية التصنيف تتضمن خمسة خطوات اساسية تبدأ بجمع صور الأشعة ، تجهيز الصور ، أستخلاص خصائص الصور ، التصنيف ،الاختبار والتقييم .أستخدمت هذه الدراسة ست خصائص مستخرجة من صور الأشعة التى أخذت من قاعدة بيانات مجتمع المعلومات ومن ثم أستناداً الي طريقة التصنيف طبقت الدراسة خوارزمية أقرب الجيران.

هذه الدراسة قسمت قاعدة البيانات الي قسمين هما: التدريب والأختبار. وبعد بناء المصنف علي أساس بيانات التدريب يتم أختبارالنموذج المقترح بإستخدام بيانات الأختبار لقياس الدقة إعتماداًعلي قيمة (k). أفضل دقة تم الحصول عليها 76.4% عند أستخدام نسبة 85% و15% على التوالى للتدريب والأختبار وقيمة K تساوي 1. وتوصى الدراسة بتحسين دقة نتيجة المصنف أستخدم خصائص أخرى قد تزيد من قوة دقة المصنف.

1.1 Introduction

Data mining is a collection of techniques to glean information from data and turn into meaningful trends and rules to improve your understanding. The basic principles of data mining are to analyze the data from different direction, categorize it and finally to summarize it .Today we are living in digital world where data increasing day by day, to get any information from mountain of database is not only difficult but mind blogging also. To deal with this huge data we need data mining technique[1].

1.2 Data Mining Definitions

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses[2].

Data mining, also known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases It is actually the process of finding the hidden information/pattern of the repositories , mainly data mining follows these steps; Data cleaning, Data integration, Data selection, data transformation, data mining, pattern evolution, knowledge evolution data reduction[3].

Data mining is a process to discover interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from large amount of data stored in databases or other information repositories[4].

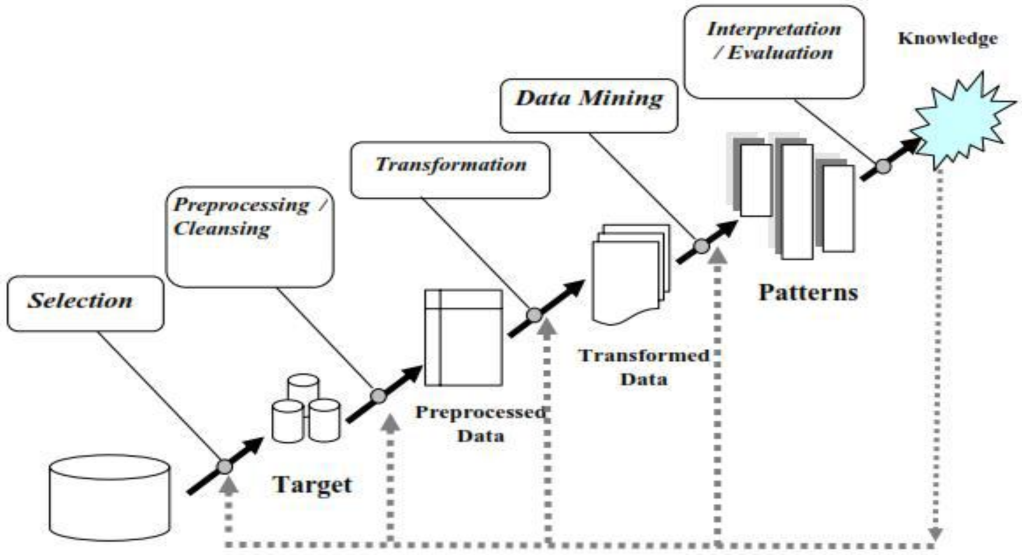

Figure (1.1): Steps of the KDD Process [3].

1.3 Data Mining Methods

Some of the popular data mining methods include decision trees and Rules, nonlinear Regression and classification methods, example-based methods, Probabilistic graphical dependency models and relational learning models. These are some famous data mining methods are broadly classified as: On-Line Analytical Processing(OLAP), Classification, Clustering, Association Rule Mining, Temporal Data mining, Time Series Analysis, Spatial Mining, Web Mining etc. These methods use different types of algorithms and data. The data source can be data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc [2] .

The most common data mining techniques are Classification, Prediction, Clustering, Association, Description, and Estimation.

1.4 Research Background

Classification is the process of finding a set of models that describe and differentiate data classes and concept , also is a group of records, any record containing member group of attributes and one of those attributes the words of class, the objective is to eventually be the records would be allocated to accurately class if possible. Divides the data set to two sections that are training set is used to build the model and test set which is used to determine the accuracy of model. Data classification is a two-step process Learning step, Where a classification model is constructed, and Classification step, Where the model is used to predict class labels for given data [5].

Medical image classification techniques is process of creating visual representations, also can play an important role in diagnostic and teaching purposes in medicine. Mammogram images is considered the most reliable method in early detection of breast cancer .Mammogram is the process of using low-dose amplitude X-rays to examine the human breast and is used as a diagnostic and a screening tool [6].

A diagnostic mammogram is used to diagnose breast disease in women who have breast symptoms or an abnormal result on a screening mammogram. Screening mammograms are used to look for breast disease in women who are asymptomatic; that is, those who appear to have no breast problems. But both screening and diagnostic mammograms depends on the radiologist accuracy reading the mammograms. with help of Computer Aided Diagnosis (CAD) breast cancer detection in mammogram images is made easier nowadays [6] .

There are other types of images rather mammogram image classification such as Magnetic resonance imaging (MRI) and Cathode Ray Tube (CRT)**,** but This research focus on mammogram image classification.

1.5 Research Problem

The accuracy of most of the classification methods depend on features extracted from the mammogram images and the classifier itself. Most of the classification algorithms or methods give only one accuracy value; this research applies classification method using K-Nearest Neighbor which can gives different classification accuracy based on different value of K.

1.6 Objectives of the Research

1. Image cropping using region of interest based on X, Y and Radius values.
2. To extract statistical features (mean, stand deviation, Skewness, Kurtosis, Contrast and smoothness) from Mammogram images.
3. To apply K nearest neighbor classification method form mammogram image classification.

1.7 Significant of Research

Classification of Mammogram helps us to detect the cancer and help the doctors to diagnosis and treatment of the disease in current situation. The computer aided diagnosis systems are necessary to assist the medical staff to achieve high efficiency and effectiveness that lead to better results in diagnosing a patient.

1.8 Research scope

This research focus on off line classification also is used mammogram image that taken From MIAS Data set. The evaluated measures that will use in this research are Confusion Matrix (true positive, true negative, False Positive and False negative) and accuracy.

1.9 Research Organization

Chapter one is a general definition about data mining and its methods and techniques Research background about general concept of classification method, medical image classification, also describe the problem statement of the research, objective, significant, and research cope of the research. Chapter two gives Literature review, Classification methods or approaches and classification methods will be use and at the end General discussion**.** Chapter three describes the research methodology, the five phases and materials and methods, dataset, Evolution measurement and Language that used in this research, Chapter four Describes the implementation of the KNN, 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 research.