Mining Health Data for Breast Cancer Diagnosis Using Machine Learning

Mohammad Ashraf Bani Ahmad

A thesis submitted for the requirements of the
Degree of Doctor of Philosophy
Faculty of Education, Science, Technology & Mathematics

December 2013

In the name of Allah, the Most Merciful, the Most Compassionate. Recite and your Lord is the most Generous - Who taught by the pen - Taught man that which he knew not. **Surat Al-`Alaq (The Clot), the Holy Quran.**

Abstract

The recent advancements in computer technologies and storage capabilities have produced an incredible amount of data and information from many sources such as social networks, online databases, and health information systems. Nowadays, many countries around the world are changing the way of implementing health care to the patients and the people by utilising the benefits of advancements in computer technologies and communications through electronic health.

Electronic health (eHealth) is the process of using emerging information and communication technologies in health care for the benefit of humans. eHealth includes a range of components such as electronic health records, electronic prescriptions, electronic and mobile treatments for patients. In Australia, the majority of medical and health coverage is provided by the government and due to shortage of medical personnel and appropriate supportive technologies, many people have to suffer long waiting times and limited medical resources. Therefore, the Australian government, territory, and state governments raised the inclusion of eHealth technologies in the health care system, to cope with the increased demand on health services and help solve some problems that face the traditional health systems.

This initiative produced the National eHealth Transition Authority Limited (known as NEHTA).The main purpose of NEHTA is to develop better ways of electronically collecting and securely exchanging health information across Australia. Since July 2012, anyone seeking healthcare in Australia can register for a personally controlled electronic health record. This can lead to a huge repository about Australian health care records.

This huge amount of data can be tuned into knowledge and more useful form of data by using computing and machine learning tools. It is believed that engineering this amount of data can aid in developing expert systems for decision support that can assist physicians in diagnosing and predicting some debilitating life threatening diseases such as breast cancer. Expert systems for decision support can reduce the cost, the waiting time, and free human experts (physicians) for more research, as well as reduce the errors and mistakes that can be made by humans due to fatigue and tiredness. However, the process of utilising health data effectively, involves many challenges such as the problem of missing features values, the curse of dimensionality due to a large number of features (attributes), and the course of actions to determine the features, that can lead to more accurate results (more accurate diagnosis). Effective machine learning tools can assist in early detection of diseases such as breast cancer, and the current work in this thesis focuses on investigating novel approaches to diagnose breast cancer based on machine learning tools, and involves development of new techniques to construct and process missing features values, investigate different feature selection methods, and how to employ them into diagnosis process.

It is believed that the adoption of electronic health systems into the health care system requires comprehensives design and development, which may need several stages to make it more useful for humans and governments. For example, storing health records and electronic exchange of health records across the country are not the only aims of eHealth. Treating health records as an important information resource and probing the data to extract useful diagnostic and disease related intelligence, by using automated approaches, including most significant features, for example, may lead to new tools/approaches to examine new cases (patients)

based on previous and similar cases, using machine learning and computer intelligence. It is the process of mapping the existing data into new unseen scenarios and settings, that can lead to increase in understanding the disease related information, such as early onset of disease, and better monitoring of different stages of disease, leading to value addition of health care technologies, for enhanced quality of service to patients, providing better assistance to doctors (bring an electronic consultant for doctors for example), and easy to cross validate standard disease diagnostic procedures. The thesis proposes several approaches to make this vision a reality. The main findings of this research can be categorised as follows:

- The thesis proposed a new approach for diagnosing breast cancer by reducing the number of features to the optimal number using the information gain algorithm, and then applies the new reduced features dataset to the Adaptive Neuro Fuzzy Inference system (ANFIS). It is found that the accuracy for the proposed approach is 98.24%, significantly better. The promising results may lead to further attempts to utilise and exploit information technology for diagnosing patients, and provide decision support to physicians.
- The thesis proposed a new approach for constructing missing features values based on iterative *k* nearest neighbours and the distance functions. The approach is an iterative approach until finding the most suitable features values that satisfy classification accuracy. The proposed approach showed improvement of 0.005 of classification accuracy on the constructed dataset than the original dataset on both Euclidean and Minkowski distance functions. The study found that Manhattan, Chebychev, and Canberra distance metrics produced lower classification accuracy on the new dataset than the original dataset. The study also noticed that classification accuracy

depends greatly on the number of neighbours (*k*). The experimental evaluation showed that less neighbours may lead to more accuracy. The reason for that, in my opinion, is the amount of noise produced from conflict neighbours. Finally, the maximum classification accuracy was on *k*=1 which was 0.9698.

- Different sets of experiments were performed to evaluate benchmark attributes selections methods on well-known publicly available dataset from UCI machine learning repository, Wisconsin Breast Cancer dataset (WBC). Naïve Bayes has performed the supreme in regard to classification accuracy. *k*-NN and Decision Tree have performed just better on dataset after applying features selection methods. In general, features selection methods can improve the performance of learning algorithms. However, no single feature selection method that best satisfy all datasets and learning algorithms.
- In regards to Classification Fusion on three well-known machine learning classifiers on breast cancer dataset. the study confirms the argument that the best combination of a set of classifiers depends on the application and on the classifier characteristics. In addition, there is no best combination of classifiers that suites all datasets. However in the current experiments, Naïve Bayes and *k*-NN produced better results when they combined as one classifier with maximum classification accuracy obtained on WBC dataset (0.9642).

Acknowledgements

I would like to thank everyone who has helped me to complete this thesis. Special, deep, and honest thanks to the supervision panel, chiefly, Dr Girija Chetty for the guidance, smile, and her advice through this research. Big thanks to all faculty staff and employees, I'm pleased for being a small part of such great place for more than three years, principally Professor Dharmendra Sharma, A/Prof. Dat Tran for his support provided through my journey, and Professor George Cho for his comments and suggestions.

My distinctive thanks to my wife; you were a complete package of family and love that gave me strength, support, and love during tough times. My son Hashim and daughter Yarra, you are my world, my words, my strength, and the reason of my life. Thank you.

Words can't express my thanks to my family, the source of power and strength; my parents for raising me up, especially my Mum, thanks aren't enough and wouldn't be enough, and sorry for being away from your warm and kind bosom but I'm back, hopefully soon. My Dad, thank you for everything you done to me, for the advice, encouragement, and support through my life. My brothers and sisters, thank you.

This thesis is dedicated to Mum, Dad, wife Fayha, son Hashim, and daughter Yarra.

Table of Contents

List of Figures

List of Tables

Acronyms

ADALINE: Adaptive linear Element

ANFIS: Fuzzy Inference System

ANN: Artificial Neural Network

CAD: Computer Aided Diagnosis

CART: Classification and Regression Tree

CES: Consistency Based Subset Evaluation

CFS: correlation based feature selection

DM: Data Mining

eHealth: Electronic Health

EHR: Electronic Health Record

ERR: Error Rate

FIS: Fuzzy Inference System

GA: Genetic Algorithm

HIS: Hybrid Intelligent System

IG: Information Gain

IGANFIS: Information Gain and Adaptive Neuro-Fuzzy Inference System

k-NN: *k* Nearest Neighbors

LSE: Least Square Estimate

MAR: Missing At Random

MCAR: Missing Completely At Random

ML: Machine learning

MNAR: Missing Not At Random

NEHTA: National Electronic Health Transition Authority

OECD: Organization for Economic Cooperation and Development

PCA: Principle Components Analysis

R: Relief

RT: Random Tree

SBFS: Sequential Backward Floating Search

SFFS: Sequential Forward Floating Search

SFS: Sequential Forward Search

SU: Symmetrical Uncertainty

UCI Machine Learning Repository: University of California Irvine Machine Learning Repository

WBC: Wisconsin Breast Cancer Dataset

WEKA: Waikato Environment for Knowledge Analysis