AN ANALYSIS OF BREAST CANCER PREDICTION USING DATA MININGTECHNIQUES

GOH E-THENG

This report is submitted in partial fulfillment of the requirements for the
Bachelor of Computer Science (Artificial Intelligence)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2010

© Universiti Teknikal Malaysia Melaka

DECLARATION

I hereby declare that this project report entitled

AN ANALYSIS OF BREAST CANCER PREDICTION USING DATA MINING TECHNIQUES

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : ______________________ Date: 28/6/2010

(GOH E-THENG)

SUPERVISOR : ______________________ Date: 30/6/2010

(DR. CHOO YUN HUOY)

© Universiti Teknikal Malaysia Melaka

DEDICATION

To my respect and love lecturers, thanks for the guidance and correction that bring me to the right way. Thanks for the advices and the challenge that given to me to let me come on with the problems and challenges.

To my dearness and beloved parents, your words and love are my greatest inspiration. You are the motivator who make me stronger and dare to face all the challenges.

To all my friends, thanks for the guidance, encouragement and support during my study life here. I really appreciate the times with all for you.

© Universiti Teknikal Malaysia Melaka

ACKNOWLEDGEMENTS

First and the foremost, I would like to express my sincere appreciation and impassion to my respected supervisor, Dr. Choo Yun Huoy, for her guidance and correction. She never feel tired and annoyed to explain and recommend solution for me. Besides, I would like to thanks to the evaluator, Dr. Burhannudin bin Mohd Aboobaider, and he always give me many idea in this project.

Due to approach the policy of University Technical Malaysia Melaka, to award the degree awarded a degree in Bachelor of Computer Science (Artificial Intelligence), students are required to complete a Final Year Project called Project Sarjana Muda (PSM). I feel grateful and thankful because I have this special chance to challenging myself to finish a PSM in the time. All the way, I learnt very much.

At last, I would like to convey my sincere to thanks for my wonderful and faithful course mates. They are willing to share their idea, knowledge and information with me. Thanks for their support. I love them very much.

© Universiti Teknikal Malaysia Melaka

ABSTRACT

Breast cancer is the number one killer decease among women in Malaysia. The rate of this decease keeps increasing for 150-200 cases every year. So, the analysis on the breast cancer is very important. This project main contribution is focus on the analysis on breast cancer prediction with data mining techniques. To fulfill and ease the analysis, the Classification Tool of Breast Cancer Dataset develops for doing the prediction of breast cancer dataset. This project consists of two (2) project methodologies which one for the experiment approach and another one for the classification tool. The classification tool has to build to contribute to the medical field especially researchers of the breast cancer datasets. User can get help with this classification tool to train and test any breast cancer dataset. This classification tool will provided three (3) data mining techniques which is Decision Tree, Naïve Bayes and Logistic to predict the outcome of the Breast Cancer dataset. The three (3) classification techniques are chosen because through a lot of literature reviews and case studies, these three (3) techniques are most suitable to predict Breast Cancer dataset and always given high accuracy of outcomes. In future, to obtain more accurate analysis on breast cancer dataset, more data mining techniques are suggested to do prediction on breast cancer dataset.

© Universiti Teknikal Malaysia Melaka

ABSTRAK

Kanser Payudara merupakan pembunuh pertama di kalangan wanita Malaysia. Kadar bagi kaum wanita menghidapi Kanser Payudara masih meningkat bagi 150-200 kes setiap tahun. Oleh itu, analisasi tentang Kanser Payudara mest dijalankan and alat untuk klasifikasi bagi data Kanser Payudara mesti dibinakan. Sumbangan muktamd dalam project ini ialah laporan analisasi bagi kanser payudara dengan klasifikasi teknik. Ini merupakan sesuatu yang sangat penting untuk sumbangan dalam kajian jangkaan bagi Kanser Payudara. Alat klasifikasi dibinakan untuk menyumbangkan sumbangan kepada bidang kanser terutamaynya Kanser Payudara. Dalam projek ini mempunyai dua (2) projek metodologi. Pertama ialah projek metodologi bagi experiment dalam jangkaan kanser payudara manakala yang satu ialah projek metodologi bagi alat klasifikasi. Pengguna boleh mendapatkan bantuan dengan alat klasifikasi untuk melatih and menguji apa-apa data Kanser Payudara.Alat klasifikasi tersebut akan memperlengkapkan dengan tiga (3) klasifikasi teknik, Decision Tree, Naïve Bayes dan Logistic untuk membuat jangkaan terhadap data Kanser Payudara. Ketiga-tiga klasifikasi teknik dipilih kerana melalui banyak kesusasteraan ulasan mendapati ketiga-tiga teknik ini merupakan teknik yang paling sesuai untuk membuat jangkaan terhapdap data Kanser Payudara. Keputusan bacaan yang dikeluarkan sentiasa memnpunyai ketepatan yang tinggi. Dalam masa depan, lebih banyak lain klasifikasi teknik ditawarkan membuat analisasi kepada kanser payudara.

© Universiti Teknikal Malaysia Melaka

TABLE OF CONTENTS

© Universiti Teknikal Malaysia Melaka

CHAPTER III PROJECT METHODOLOGY

CHAPTER IV TECHNIQUES

CHAPTER V CLASSIFICATION TOOL FOR BREAST CANCER DATASET ANALYSIS AND DESIGN

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

LIST OF ATTACHMENTS

© Universiti Teknikal Malaysia Melaka

CHAPTER I

INTRODUCTION

1.1 Projects Overview

This project proposed with theory of data mining in various classification methods to do analysis and comparison different classifier on Breast Cancer dataset. Besides, to accomplish the analysis and comparison, develop a *classification tool.* This tool allows users to load various kinds of breast cancer datasets to predict the accuracy of the dataset based on AI (Artificial Intelligence) technique. In this project, breast cancer data will be chosen as a testing data and test run with three (3) data mining classification method which is Decision Tree, Naïve Bayes and Logistic. The data mining classification algorithm will be written in .java language to predict the breast cancer datasets. This idea allow the classification tool to run the algorithms of various types of data mining methods and to justify which is more suitable and accurate based on the input dataset given.

This project will be run with java programming which will provide an interface for user to input the dataset and view their output. The result will be called and show to the user with call function method. This classification tool's main purpose is to compare the several of the classification method with various types of breast cancer data and determine which classification method is the most suitable method to determine the accuracy.

Classification is a tool for understanding relationships of living things. With classification, similar things will be group together. In this project, a classification is development with various fuzzy methods to classify the breast cancer data. Output is

© Universiti Teknikal Malaysia Melaka

how accuracy this method works on this dataset. The dataset will be split to run few times to see the different of the output. To further understanding of the technique to the dataset, comparison and analysis must be taken based on the output result. A higher accuracy gets meant that the classification method is the most suitable method for the dataset.

1.1.1 Project Background

Now a day, many represent classification tool can use to predict various kinds of datasets. Sometimes, the method provided by the represent classification tool is too general and not enough specific to predict breast cancer data. So, from a lot of research paper and study, the three (3) best classification methods are selected to build a classification tool to predict breast cancer datasets. The best techniques were a) Decision Tree, b) Naïve Bayes and c) Logistic. This tool is helpful to help the patient with the problem with breast cancer. Since, breast cancer is a common form of cancer affecting women in Malaysia. A late diagnosis will cause the women's life to be risky. Therefore, these kinds of classification tools are important to develop for contribution in the prediction of breast cancer dataset.

1.2 Problem Statement

In the analysis of breast cancer dataset, there is no any definite guidance or contribution of suitable classifier for Breast Cancer dataset. Furthermore, there is no specified classification tool for prediction of Breast Cancer dataset. Research and result analysis should doing to avoid the random selected classification techniques to do prediction will cause or produce an unsure result.

© Universiti Teknikal Malaysia Melaka

1.3 Objectives

1. To propose classification methods for breast cancer datasets.
2. To build a classification tool for breast cancer dataset
3. To compare and analyze the results between bench mark data mining techniques based on selected breast cancer dataset.

1.4 Project Scope

This project may focus in using Decision Tree, Naïve Bayes and Logistic data mining techniques to analyst and classify the breast cancer dataset. These data mining techniques are chosen as this classification tool because many researchers and literature studies had been made to those techniques. The classification tool able to accept .csv and .arff data file. The breast cancer datasets may in the form of numerical and nominal.

The breast cancer dataset predict using percentage split. The dataset will be split size of 10 with different training set and testing size. The parameters of these tools depend on the accuracy of the techniques on breast cancer dataset. The target user may be any users who want to do prediction on Breast Cancer dataset.

1.5 Project Significance

This project is important for the breast cancer dataset researches. All the 3 breast cancer dataset prediction result is based on the accuracy and AUC (area under curve). All the result is recorded and does comparison to get the most suitable technique on breast cancer dataset.

The classification tool is important for the users to train data to generate the rule which can classify the data into several clusters. Development of this classification

© Universiti Teknikal Malaysia Melaka

tool capable user to predict any medical datasets therefore several rules and accuracy will be an output depends of the training data of the user. Different data will have different rule and accuracy. This classification tool is a combination of various fuzzy methods to easier user to train their data in an interface. Then, user enables to save the output result as text. By this way, users save their time for thinking to choose which classification method to predict the medical data because these projects' main purpose is to build a highest accuracy classification tool based on medical datasets especially breast cancer data. After train with various kinds of data, user may get the pattern of the classification method. User will get the result of which method is suitable for numerical datasets and which method is suitable for nominal datasets. With this pattern, user will save their time with training other dataset next time. So, training more data with 1 classification method will be generating more accuracy rule.

1.6 Expected Output

In this project, the output will be an analysis researches about the breast cancer dataset. The report may be depends on the percentage of accuracy based on the breast cancer datasets which the users want to predict. The result show in 10-cross-validation due to the value of training filed and testing filed is different, then the output result and the accuracy may be different. Besides, user can choose only to get the average result.

Due to accomplish the researches of the breast cancer dataset, a classification tool shall be build to predict the breast cancer dataset result with coded with data mining techniques.

© Universiti Teknikal Malaysia Melaka

1.7 Conclusion

As conclusion, in the end of this project, a classification tool will be build by using three (3) popular data mining techniques which is Decision Tree from Machine Learning and Naïve Bayes and Logistic from Statistic. This project requires breast cancer datasets to complete the prediction. To build a highest accuracy of classification tool to predict these breast cancer datasets, a large dataset is acquired. The best datasets must undergo long process of data collection, cleaning and transformation before classify.

Although the survivability of breast cancer is highest for early stage patient but this disease is very common among women. So, a high accuracy of classification tool based on breast cancer datasets must be well-formed to benefit the entire user and contribute in medical field.

© Universiti Teknikal Malaysia Melaka

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

This chapter will discuss about the methods use for classification tool and carry out the preliminary study and literature review about the project. Basically, this chapter includes facts and finding, project methodology, project requirement, project schedule and milestone. This chapter's objective in this study is to make sure the gather information is useful for this project and meet project and user requirement.

2.2 Literature review

A literature review is a body of text that aims to review the critical points of current knowledge and or methodological on a particular topic. In this chapter, many Artificial Intelligence technique like Decision Trees, Naïve Bayes and Logistic in Breast Cancer datasets will be researched and understand to help in complete this project.

2.2.1 Breast Cancer in Malaysia

Breast cancer is the most common cancer among Malaysian women. There is a marked geographical difference in the worldwide incidence of breast cancer, with a higher incidence in developed countries compared to developing countries (Hishan

© Universiti Teknikal Malaysia Melaka

AN, Yip CH, 2004). The breast clinic in Kuala Lumpur Hospital diagnosed approximately 150 to 200 new cases of breast cancer a year. Although it appears that the incidence of breast cancer in Malaysia is lower than in the developed countries, the difference may be attributable to the difficulty in getting accurate statistics and to underreporting of cases.

Nonetheless, from the available data, it is clear that breast cancer continues to be the most common cancer among Malaysian women. The strongly negative social-cultural perception of the disease, made worse by the geographical isolation of many rural areas, accounts for the delayed diagnosis and the often advanced stage of disease at presentation. A prospective population-based study is called for to verify the demographic patterns of breast cancer, particular in Malaysia and other developing countries. The findings of such a study may have implications for future breast screening programs and for facilitating the understanding of differing risks of breast cancer among women around the world.

The National Cancer Institute (NCI) is funding numerous research projects to improve conventional mammography (an x-ray technique to visualize the internal structure of the breast) and develop other imaging technologies to detect, diagnose, and characterize breast tumors. But, no matter how high technology help in detection of breast cancer, the pattern of a symptom of breast cancer is really hard to define. So, the decease of breast cancer must be convert and record in a database as a dataset to doing further research and classification.

2.2.2 Classification Techniques

This study evaluates the performance of classification techniques with the application of a new single classification tool to predict the breast cancer datasets. The classification technique has been tested on three (3) breast cancer datasets. The study will help researchers to select the best suitable technique of classification problem for breast cancer datasets in term of classification accuracy.

© Universiti Teknikal Malaysia Melaka

However, many others use other data mining techniques to predict Breast Cancer dataset. At year 2007, Han-Yu Chuang and friends used network-based as data mining technique to classify breast cancer metastasis. They found that sub-network markers are more reproducible than individual marker genes selected without network information, and it achieve higher accuracy in the classification of metastatic versus non-metastatic tumors. A protein network-based is applied to approach that identifies markers are not as individual genes but as sub-networks extracted from protein interaction databases.

Besides, at year 2003, Ravi. Jain and Ajith. Abraham doing a experiment to examine the performance of four fuzzy rule generation methods on Wisconsin breast cancer data. They generate fuzzy rules to specify the membership function of each antecedent fuzzy set using the information about attribute values of training patterns. The performance of each approach is evaluated on breast cancer data sets. In the end, the simulation results show that the Modified grid approach has a high classification rate of 99.73 %.

2.2.2.1 Decision Tree

According to Nguwi Yok Yen (2006), Decision Tree is a powerful and popular tool for classification and prediction. It attempts to model the structure of a tree. In decision tree classifier, we have a root, and start branching out to more nodes. If a branch has too few instances, we prune it to prevent it from further developing. Decision tree classify instances by sorting them down the tree from root node to decision node, which provides the classification of the instance.

Besides, Decision tree also defined as an alternative choice available at each stage of deciding how to manage a clinical problem, display graphically; at each branch or decision node, the probabilities of each outcome that can be predicted are shown; the relative worth of each outcome is described in tern of its utility or quality of life (Fahad Shahbaz Khan, 2008).

© Universiti Teknikal Malaysia Melaka

According to the Alaa M. Elsayad (2010), Decision Tree is a good and suitable technique on Breast Cancer Dataset. With the helpful tool, support vector machine (SVM), a complex decision boundary able to create to classify the good ability between two classes. Alaa M. Elsayad used many nodes as a tool to help in analysis on Breast Cancer dataset. As the conclusion, he found that Decision Tree identifies only five attributes to get 98.95% accuracy on training subset and 100% accuracy on the test subset.

According to Radim Belohlavek (2009), selection using Decision Tree with formal analysis concept (FCA) found out interconnects areas of decision trees and formal concept analysis and makes comparison between two (2) different Decision Tree algorithms. The further researches will based on explore the possibility to compute a smaller number of formal concepts from which the nodes of a decision tree is constructed.

2.2.2.2 Naïve Bayes

According to Harry Zhang (2004), Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. A **Naive Bayes classifier** is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without believing in Bayesian Probability or using any Bayesian methods. Naive Bayes can be used for both binary and multiclass classification problems.

$$p(A|B)=\frac{p(B\cap A)}{P(A)} \text{ is a formula of Bayes Theorem} \qquad (2.1)$$

© Universiti Teknikal Malaysia Melaka

According to Newton Cheung (2001) researches, the Bayesian Network is using to do experiment on medical database collected from the University of California. The performance measurement will be based on the sensitivity, specificity and computational time. Those performance measurements will be compared with Decision Tree. On the experimental of Newton Cheung, he found that the Bayesian Network performance better than Decision Tree and Naïve Bayes because the Bayesian Network showed the competiveness with the Naïve Bayes classifier. However, the Naïve Bayes classifier has the computational advantages than all the Bayesian Network classifier.

2.2.2.3 Logistic

In statistics, Logistic Regression is a model used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. It makes use of several predictor variables that may be either numerical or categories. For example, the probability that a person has a heart attack within a specified time period might be predicted from knowledge of the person's age, sex and body mass index. Logistic regression is used extensively in the medical and social sciences as well as marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription.

According to Dursun Delen (2004), he used logistic regression as a classifier to predict the breast cancer survivability. Logistic regression is a generalization of linear regression so it enable used for predicting multi-class and binary dependent variable. Logistic regression able to build the model

2.3 Fact and finding

A literature review is very important in this section. Go through the research, many knowledge and experience will be learn and help in this project. For more information, some case studies to techniques developed by some researchers have been given as below:-

© Universiti Teknikal Malaysia Melaka

2.3.1 Case study 1 - Predicting breast cancer survivability: a comparison of three data mining methods

Based on Dursun Delen' , Glenn Walker and Amit Kadam (2004), Breast cancer is a major cause of concern in the United States today. Though predominantly in women, breast cancer can also occur in men. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. With the increased use of computers powered with automated tools, storage and retrieval of large volumes of medical data are being collected and are being made available to the medical research community who has been interested in developing prediction models for survivability.

Furthermore, medical applications of data mining include prediction of the effectiveness of surgical procedures, medical tests and medications, and discovery of relationships among clinical and pathological data. Naïve Bayes is a special form of Bayesian network has been widely used for data classification in that its predictive performance is competitive with state-of-theart classifiers such as C4.5 (Ranjit Abraham, Jay B.Simha, Iyengar S.S,2007).

Many new techniques are developing due to increasing of the volume of data. A major area of development is called KDD (knowledge discovery in databases). KDD encompasses variety of statistical analysis, pattern recognition and machine learning techniques. KDD is a formal process whereby the steps of understanding the domain, understanding the data, data preparation, gathering and formulating knowledge from pattern extraction, and "post-processing of the knowledge" are employed to exploit the knowledge from large amount of recorded data. The step of gathering and formulating knowledge from data using pattern extraction methods is commonly referred to as data mining. Applications of data mining have already been proven to provide benefits to many areas of medicine, including diagnosis, prognosis and treatment.

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5, and

© Universiti Teknikal Malaysia Melaka

Breiman et al.'s CART . As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms (e.g., information gain, Gini index, and Chi-squared test) to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID). Based on the favorable prediction results we have obtained from the preliminary runs, in this study we chose to use C5 algorithm as our decision tree method, which is an improved version of C4.5 and ID3 algorithms.

In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, sensitivity and specificity). The results were achieved using 10 fold cross-validation for each model, and are based on the average results obtained from the test dataset (the 10th fold) for each fold. However, the decision tree (C5) preformed the best of the three models evaluated. The decision tree (C5) achieved a classification accuracy of 0.9362 with a sensitivity of 0.9602 and a specificity of 0.9066. The ANN model achieved a classification accuracy of 0.9121 with a sensitivity of 0.9437 and a specificity of 0.8748. The logistic regression model achieved a classification accuracy of 0.8920 with a sensitivity of 0.9017 and a specificity of 0.8786.

Several issues involved with the data collection, data mining and the predictive models that warrant for further discussion:-

- Amount and quality of the data. Medical databases may consist of a large volume of heterogeneous data, including heterogeneous data fields. Additionally, as with any large database, medical databases contain missing values that must be dealt with prior to the use of the data mining tools.
- Data mining has been criticized by some for not following all of the requirements of classical statistics. For example, most data mining tools use training and testing sets drawn from the same sample. Under classical

© Universiti Teknikal Malaysia Melaka

statistics, it can be argued that the testing set used in this instance is not truly independent, and therefore, the results are biased.

- Data mining can provide useful information and support to the medical staff by identifying patterns that may not be readily apparent, there are limitations to what data mining can do.

2.3.2 Case study 2 - Wisconsin Breast Cancer Data with Decision Tree

According to Nguwi Yok Yen (2006), Decision tree is a powerful and popular tool for classification and prediction. It attempts to model the structure of a tree. In decision tree classifier, we have a root, and start branching out to more nodes. If a branch has too few instances, we prune it to prevent it from further developing. Decision tree classify instances by sorting them down the tree from root node to decision node, which provides the classification of the instance. Each node of a tree is either:

• a *leaf node* - indicates the value of the target attribute, or

• a *decision node* - specifies some test to be carried out on a single attribute, with one branch and sub-tree for each possible outcome of the test.

In this paper, the researcher use the DTREG as an analytical tool that builds classification and regression decision trees, support vector machine (SVM), discriminate analysis and logistic regression models to describe data relationships and can be used to predict values for future observations. A decision tree in DTREG is represented as a binary (two-way split) tree that shows how the value of a *target variable* can be predicted by using the values of a set of *predictor variables*. For example, in the tree shown below, nodes 2 and 3 were formed by splitting node 1 on the predictor variable "Uniformity of Cell Size". Cell size of 1 and 2 belongs to node 2, and the rest belongs to node 3.

© Universiti Teknikal Malaysia Melaka