CLASSIFICATION OF MICROARRAY DATASETS USING RANDOM FOREST

NG EE LING

UNIVERSITI SAINS MALAYSIA
2009

CLASSIFICATION OF MICROARRAY DATASETS USING RANDOM FOREST

By

NG EE LING

Thesis submitted in fulfillment of the requirements
for the degree of
Master of Science

June 2009

Acknowledgement

Many individuals have contributed to the success of this research. These individuals consist of my supervisor, Dr. Yahya Abu Hasan, my family and friends.

I would like to express my deepest gratitude to Dr. Yahya Abu Hasan, who is the supervisor of this master's research for his continuous support. His bright advice and constructive ideas have become the main factor towards the success of this research. I also want to thank him for sharing with me the important writing techniques and for being so patient with me throughout the whole process.

Besides that, I would like to express gratitude to the Institute of Graduate Studies and School of Mathematical Sciences for granting me the entitlement of fellowship for a total of three semesters. The financial support obtained had been of great help in my studies.

I also want to thank my fellow post-graduate mates and friends who have shared their ideas with me.

Not to forget, I am grateful to my family who have consistently supported me. Their support has indeed boosted my confidence.

Lastly, I thank God for making everything possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

KLASIFIKASI SET DATA TATASUSUNAN MIKRO MENGGUNAKAN RANDOM FOREST

ABSTRAK

Teknologi DNA tatasusunan mempunyai keupayaan untuk memerhati lebih daripada ribuan nilai ekspresi gen dalam satu chip. Ia juga mendatangkan kebaikan dalam bidang perubatan kerana ia dapat membantu dalam pengesanan mutasi genetik dan penyakit. Kewujudan satu model yang baik dapat meramalkan kelas penyakit yang tidak diketahui sebelumnya. Untuk mendapatkan satu model yang baik, kita mesti terlebih dahulu mmperolehi keputusan klasifikasi yang baik. Namun, kebanyakan data tatasusunan mempunyai bilangan gen yang melebihi bilangan sampel. Oleh itu, untuk mendapatkan keputusan klasifikasi yang baik, bukan sahaja pemilihan jenis klasifikasi penting tetapi juga ciri penting dalam gen yang dipilih. Dalam penyelidikan ini, kita telah mencadangkan satu cara dinamakan 'stair-line' dalam pemilihan gen yang penting untuk mengurangkan kesan kurtosis yang wujud. Klasifikasi yang digunakan ialah 'Random Forest'. Lima set data tatasusunan dengan bilangan gen dan sampel yang berlainan telah digunakan untuk mempamerkan keupayaan cara 'stair-line' yang dicadangkan. Cadangan ini telah memperbaiki peratusan kebetulan dalam keputusan klasifikasi dan pada masa yang sama telah mengurangkan kesan kurtosis yang wujud dalam gen. Selain itu, pengklasifikasi yang lain juga telah dipertimbangkan dan keputusan yang diperolehi telah dibandingkan dengan keputusan yang diperolehi dengan menggunakan 'Random Forest'. Secara kesuluruhan, keputusan yang diperolehi dengan menggunakan Random Forest adalah lebih baik jika dibandingkan dengan keputusan yang diperolehi dengan menggunakan klasifikasi lain.

CLASSIFICATION OF MICROARRAY DATASETS USING RANDOM FOREST

ABSTRACT

DNA microarray technology has enabled the capability to monitor the expressions of tens of thousands of genes in a biological sample on a single chip. Medical fields can benefit from microarray data mining as it helps in early detection of genes mutation and diagnosis of disease. A well built model can be used to predict unknown disease classes in a test case. Prior to a well built model is to achieve good classification results which rely very much on the classifiers that are being used. However, in most microarray data, the number of genes usually outnumbers the number of samples. Thus, it is often not just selecting the type of classifier that is essential but also the features looked in selecting significant genes that will contribute to good classification results. Genes selection also varies from study scope and depends on the criteria researchers are looking at. In this study, we propose a stair-line method to select significant genes to reduce the effect of kurtosis found among the genes. Classification is then done using Random Forest. Five microarray datasets with different number of genes and samples are used to demonstrate the effectiveness of this method. This method improves the percentages of correct classification and at the same time reduces the effect of kurtosis in the genes expression values. Other conventional classification schemes are also looked at as a comparison to Random Forest and it is shown that the latter is one classifier that is more superior to the others. In short, Random Forest managed to give a competitive result in classifying genes correctly as Random Forest performed consistently well on all datasets.

CHAPTER 1

INTRODUCTION

1.1 Knowledge Discovery in Database

Knowledge discovery in databases (KDD) is the analysis of data. It is the practice of sorting through data to identify pattern and to establish relationship. The main reason in doing so is to discover previously unknown information that might be potentially useful in the future. With the availability of advanced mining tools which use artificial intelligence, statistical methods or pattern recognition plus the availability of the abundance of data, people are able to perform data mining on various sequences to achieve various outcomes.

There are various methods in which one can adopt to perform data mining. These methods are often recognized as the data mining parameters. Some of the often used methods which include the following:

Association - looks for patterns where one event is connected to another event

Sequence or path analysis – looks for patterns where one event leads to another later event

Classification – finding a model that describes data classes so as to use the model for future prediction

Clustering – finds and visually documents groups of fact that is not previously known

Predictions or Forecasting – discovers patterns that can lead to predictions about the future (Olson and Delen, 2008)

1.2 Microarray Data Mining

Genomic study has been of great interest over the past years. Genomic study involves gene analysis tasks which are carried out to identify and learn characteristics of genes which can lead to many hidden potential information. One of the potential information that has been looked into by the bioinformatics community is the identification of diseases. In the past, genomic study was done by looking at one gene at a time. This technique is not only tedious but also has a potential of lack of information because it is only capable of generating limited results and at a time. Now, with the advancement of microarray technology, this can be done very easily.

Microarray technology has given researchers the opportunity to perform genomic study by looking at thousands of genes simultaneously instead of just one gene at a time. This technology enables the measurement of tens and thousands of gene expressions of a biological sample in just one single chip (Samb, 2005).

Microarray data usually consists of two sections, the samples and variables or genes. Measuring gene expression using microarray is relevant to many areas of biology and medicine. The uses of microarray in the field of medicine vary and they include DNA microarray, tissue microarray, protein microarray, plant microarray and many more which adds to the reasons why microarray data is mined so widely since its existence.

Microarray data mining is indeed a very useful study as it helps in early detection of genes mutation and diagnosis of disease of which, if diagnosed early can help prevent death. Hence, microarray data mining which uses the combination of both mathematical modeling and biological technology is certainly a comprehensive way not only to classify disease but also to examine disease outcome and discover new cancer subtypes. Some recognize this field as the field of Bioinformatics.

Cancer classification has been a popular study over the past few years. Just like any other data, cancer too come in different subtypes. In classification problems, these subtypes are known as classes. Classification can therefore be done onto cancer data to build a model that can describe the classes. Previously, cancer classification is done using the most traditional method which is based on combinations of few clinical techniques. These techniques include looking at the differences of the cell shapes and detecting enzymes that are not normally produced by certain cells. The former are the clinical methods that are carried out to help diagnose cancer disease. However, studies show that not one of those tests are 100% accurate and are always inconclusive (Twyman, 2002).

Just like when mining other types of data, many challenges are faced when mining microarray data. Microarray data is one data which contains the expression levels of thousands of genes, thus increasing the difficulty level when it comes to mining the data. Secondly, microarray data usually has a very large number of variables as compared to the observed samples. And therefore efforts to achieve good results very much depend on the study scope of the researcher. Some researchers might classify good

results as obtaining good models whereby they obtain high percentage of correct classification while some chose to look at the error rates of models obtained instead. For example, Ng and Breiman (2005) decided to use Random Forest to select their first 20 important genes before they used their proposed bivariate selection method to see the interaction among genes and further reduce the number of genes to obtain better results. Nevertheless, although mining microarray data might be a wearisome task, the result obtained is often worth the effort.

1.3 Objective

Classification, which allows us to find a model that describes data classes, is the main mining method in this research. Classification not only allows us to classify genes but also to see hidden patterns especially among significant genes in order to obtain better results. Most classification schemes rely very much on selecting useful or important genes which can contribute significantly to the classification results and thus creating a good model.

The main objective of this research is to come out with good classification accuracy (high percentage of correct classification of genes) by identifying smaller set of genes. We propose a stair-line method to select significant genes. Basically, our stair-line method involves three steps which consist of first selecting significant genes using Random Forest, second eliminating genes with odd platykurtic behaviour and third re-select top 20 significant genes with highest t-values. Here, we define significant genes as those genes which are well differentially expressed. Lastly, classification is then done

using Random Forest classifier of which its error function has been modified to reduce the effect of kurtosis.

1.4 Methodology

Data mining tasks vary from one study to another. The fundamental stages that are involved in data mining include pre-processing, processing and post-processing. The initial data mining task in our research involves selecting significant genes to reduce the effect of kurtosis found among the genes. While many other researchers have chosen to work at specific algorithms, we have chosen to look at the effect of the statistical measurement kurtosis instead as this is an area which has not much been emphasized on. Teschendorff et. al. (2006) used kurtosis behaviour found among genes as a clustering method.

In our paper, we have proposed the stair-method which consists of a total of three steps in selecting significant genes before classification is done to reduce the kurtosis effect found among genes. While we could have only used Random Forest classifier to select important genes, we want to bring in the importance of distribution in genes and show that the selection of significant genes does not necessarily involve only one or two steps but three as shown in our research. However, we have also proven that the three steps chosen synchronized well with each other, giving good and reasonable results.

Once a raw data has been obtained, it must go through certain steps of data cleaning before it can be processed. Thus, the data cleaning process, often referred to as

pre-processing stage is absolutely vital as it is the initial stage to start the whole data mining processes. As microarray is normally a large dimensioned data, pre-processing is usually not an easy task. In our study, we have introduced a stair-line method to choose the significant genes. This stair-line method will be done using scripts written in Mathematica, a computer algebraic system.

Once a raw data has been pre-processed or cleaned, mining methods can be applied onto it. As mentioned earlier, the mining method used for this research is classification. Besides using the Random Forest classifier, the use of a powerful data mining software, WEKA (Waikato Environment for Knowledge Analysis), and the readily available several classification algorithms in WEKA will also be used to build our classification models. These algorithms include J48, ZeroR, k-nearest neighbour, Naïve Bayes, support vector machine and neural network and will be used as a comparison to the Random Forest classifier.

1.5 Summary of Contribution

In this research, we have proposed an alternative method to select significant genes, which is by looking at the genes' distribution. Normal procedure of selecting significant genes usually involves only one or two steps. Tibshirani et. al. (2002) who has created an approach known as nearest shrunken centroids to identify subsets of genes that best characterize each class in classifying the blue-cell tumor. However, their research was only validated by the blue-cell tumor and leukemia dataset which could possibly mean that the method might not work well for the other cancer datasets. An

example on research on the usage of two algorithms was done by Li et. al. (2002). In their paper, a Bayesian method which performed similarly to that of support vector machine's algorithm was used. Nevertheless, they also limited their research to only three datasets and have also not clearly proven their Bayesian method's superiority over the other methods.

Our proposed stair-line method has three steps and all these three steps synchronize well with each other. We have also opted to use five datasets instead to show the superiority of our proposed method. Our methods deviate from the conventional way of selecting significant genes. Our combination of steps looked at both genes' distribution as well as how the genes are differentially expressed. Initial experiment showed that these genes generally have a negative kurtosis value or are of platykurtic distribution although there are some outliers. After omitting the outliers and selecting genes which are differentially expressed using a t-test, we reduce the effect of kurtosis by modifying the error function in the Random Forest classifier. Our study has also successfully proven that Random Forest is a versatile classifier yet robust enough to handle highly dimensioned data such as the microarray data.

1.6 Thesis Summary

This thesis has six chapters. It starts with the introduction chapter which gives a brief but precise explanation on microarray data mining and its issues that motivate this research. The introduction also tells the study scope and the layout of our research.

In chapter two, we highlight the importance of data mining and different methods of data mining. Besides, the statistical measurement and the classification techniques used in this study are also explained. We also discuss about the other classification methods which are the J48, ZeroR, k-nearest neighbour, Naïve Bayes, support vector machine and neural network that are being used in this study as a comparison to our main classifier, Random Forest.

Chapter three introduces the technology of microarray. This chapter focuses on introducing the fundamental of microarray including the process and its connection with human genes. We also highlight the common issues faced in microarray data mining and past researches that have been done in this field.

The following chapter which is chapter four discusses the implementation of our experiments. Here, we introduce our datasets in detail and explain how our data is being prepared using our proposed method which is the stair-line method before classification is done.

Chapter five presents and discusses our results. Results are discussed in details and graphs and tables are used to show a better representation of the results.

The last chapter is the conclusion of the thesis. In this chapter, we recapture the purpose of this study as well as our objective and motivation.

CHAPTER 2

DATA MINING AND TECHNIQUES OF CLASSIFICATION

2.1 Data

2.1.1 Collection of Data

Data and information of different forms are created and stored each day. These data are collected for a variety of reasons. We are indeed overwhelmed with the amount of data in the world, and this amount seems to be increasing with no end in sight. Some data are so huge that it requires computers with larger memory capacity to handle them. Hence, it is almost impossible to imagine having such data handled manually without the help of computer technology.

The phenomenon of data-handling is actually closely related to the development of the computer technology. Computers have now made it easy to save and store information. There are a lot of advanced tools that are available to store data. Examples of such tools that are available are Structured Query Language (SQL) and Oracle or Microsoft Access. With the availability of these tools, we can store whatever data we want in a clearer form with the additional benefit that this data can be retrieved anytime and anywhere and in a much convenient way.

However, while having to store data efficiently is important, good human skills are essential when it comes to understanding the data that is being stored. Most of the time, people tend to lack the skill in understanding the data collected and might eventually not be able to interpret the collected data properly. As there might be hidden information in the data that can be potentially useful, the former is

definitely a serious problem. Therefore, knowledge discovery is introduced in the hope to solve this and other matters arising that are connected closely to data-understanding.

2.1.2 Data and Its Quality

There are many forms of data and they usually come in different dimensions. While some expects a large data to contribute to more new findings, it also requires far more complicated methods to handle the data.

Microarray data for example, is one data that is very large in dimension and contains more number of variables (genes) as compared to the number of samples or observations. Hence, carrying out analysis on the data is definitely going to require more time and effort.

Besides, it is vital to have an overview of the type of data we are mining. To do this, the data's pattern must be evaluated so as to obtain a clearer picture of the data which can then enhance the process of data mining.

Looking at the data's pattern involves the usage of statistics. Spiegel (1999) mentioned that statistics is a scientific method relating to the collection, analysis, summarization, and explanation of data. There are many ways to look at the pattern of a data which include investigating on the measures of central tendency which involve the calculation of mean, median and mode and measures of dispersion which involve the calculation of first quartile, third quartile, variance and standard deviation. Some also consider the data's maximum and minimum values to help

locate any outliers in the data. Missing values is another problem that is commonly faced especially when handling real-life data. Depending on which variables these values are missing on, researchers will conduct necessary steps, either to substitute the missing values with a certain average point or to disregard the whole variable This again depends on the researchers' scope of study.

Unlike the commonly used statistical measurement like the measure of central tendency and measure of dispersion, the measurement of kurtosis is one criterion we looked at in this study.

Kurtosis is the degree of peakedness of a distribution. Mathematically, kurtosis is the normalized form of the fourth order central moment of a distribution. A high kurtosis value means a higher variance which is due to the infrequent extreme deviations, as opposed to the frequent modestly sized deviation. Kurtosis is useful to characterize the characteristics of a distribution (Pearson, 2005). Unlike skewness which can be easily seen from a box plot, kurtosis is often not as easily detected. Nevertheless, kurtosis can be calculated by using $\frac{\sum(x-\mu)^4}{N\sigma^4}-3$ whereby, x is the value of a point, μ represents average and σ represents standard deviation of the data. An approximate standard error to compensate the existence of non-zero kurtosis is given by $e=\sqrt{\frac{24}{n}}$ (Crawley, 2005).

2.2 Data Mining

Data mining has become a powerful technology in different fields. The term data mining was used to describe the component of the Knowledge Discovery in Databases (KDD) process where the learning algorithms were applied to the data and can be defined as the process of selection, exploration and modeling of large quantities of data to discover models and unknown patterns (Giudici, 2003).

Data mining is a whole process of data extraction and analysis to achieve specified goals. Data mining is different from data retrieval because it looks for relations between phenomena that are not known beforehand. So, in short, data mining is about solving problems by analyzing data that is already present in the databases (Olson and Delen, 2008).

Data mining uses techniques such as artificial intelligence, statistics and pattern recognition. Data mining methodologies include:

Association - looking for patterns where one event is connected to another event

Sequence or path analysis - looking for patterns where one event leads to another later event

Classification - looking for new patterns

Clustering - finding and visually documenting groups of facts not previously known

Forecasting - discovering patterns in data that can lead to reasonable predictions about the future.

A complete data mining process comes in three steps which are the pre-processing, processing and the post-processing. The pre-processing step is often

known as the feature selection step whereby researchers reduce the number of variables by getting rid of noisy and irrelevant ones. Clustering and classification are types of processing method where the former is unsupervised and the latter is supervised. Forecasting on the other hand is a post-processing task.

While there are many data mining methodologies available, classification is probably the oldest and most widely-used of all when it comes to mining microarray data and will be used throughout our study. There are a few classification techniques which will be used in this study apart from our main concern which is the Random Forest. Those classification techniques are ZeroR, J48, Naïve Bayes (NB), k-nearest neighbour (KNN), support vector machine (SMO) and neural network (MLP).

2.3 Classification and Its Techniques

2.3.1 Random Forest

Random Forest is an algorithm that is able to compute a collection of single classification trees. Random Forest is a classification algorithm developed by the late Leo Breiman in 2001.

Random Forest creates a forest-like classification. The basic algorithm in Random Forest works in such a way that each tree is constructed using a different bootstrap sample built from the original data. The each tree that is built is grown to the fullest without any pruning. The bootstrap data points is a random sample of size n drawn with replacement from the sample (x_1, ..., x_n). This means that the bootstrap data set consists of members of the original data set, some appearing zero times, some appearing once twice, etc (Efron and Tibshirani, 1997). The whole bootstrap

procedure is repeated several times, with different replacement samples for the training set and the result is then averaged.

The bootstrap sample usually consists of about two-thirds of the data. The other one-third or out-of-bag (oob) case will then be used as the 'test' set to get the classification result. Classification is done by getting the majority vote (particular class) of each 'test' set in a certain collection (Breiman, 2001).

Random Forest has its own variable (genes) selection procedure. The number of votes cast for the correct class is counted after each out-of-bag case is put down in each tree grown in the forest. The values of the mth variable in the oob cases are then permuted and put down the trees. The difference between the correct votes cast for the variable-permuted data and the untouched data is calculated by subtracting the former from the latter. The raw importance score for the mth variable is the average over all trees in the forest.

Random Forest is a good classifier because it gives competitive results in accuracy among current algorithms. Besides, it has the capacity to run efficiently on large data which means it can handle thousands of input variables and this is definitely an important feature in our study as we dealt with microarray data which contains thousands of variables.

2.3.2 Decision Tree (J48)

Decision tree is derived from the simple divide-and-conquer algorithm. The most common algorithms of the decision trees are C4.5 and ID3. The attractiveness

of decision tree is its easy and convenient representation whether in visualization or in rules that can readily be expressed so that human can understand them (Gamberger et. al., 2001).

J48 is one classifier that is implemented based on the concept of decision tree and uses the C4.5 algorithm. It generates pruned and un-pruned C4.5 algorithm decision tree. C4.5 allows pruning of the resulting decision tree. Although pruning tends to increase the error rates on the training data, more importantly, it can decrease the error rates on the unseen test cases (Witten and Frank, 2000).

The decision tree algorithm works by first selecting an attribute to be the root node and make a branch for each possible value. So, this will split the instances into subsets. When all instances at a node have the same classification, the tree will stop splitting. In short, the decision tree is a classifier that works in the form of a tree structure (Gamberger et. al., 2001). Figure 2.1 shows a visualization of a tree structure classifier.

On the whole, J48 can be considered as a good classifier as it is able to deal with numeric attributes, missing values and noisy data. Nevertheless, the drawback is that only one attribute is used to split the data into subsets at each node of the tree. Besides that, J48 usually only performs better with binary-class data as compared to multi-class data.

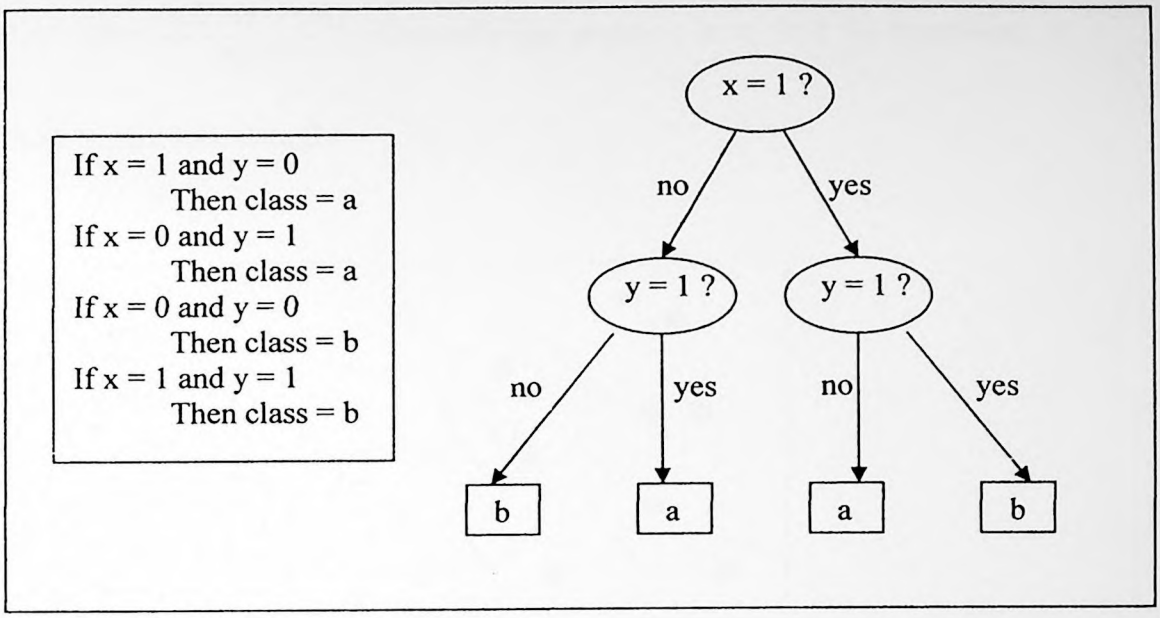

Figure 2.1: Visualization of a Tree

2.3.3 Bayesian Theorem - Naïve Bayes

The Naïve Bayes Classifier technique is based on Bayesian theorem. The Naïve Bayes classifier has been successfully applied in a number of machine learning applications. It is constructed by using the training data to estimate the probability of each class given the gene expression of the new sample. The Naïve Bayes model makes additional assumption that the values for each attributes are independent (Aas, 2001).

Naïve Bayes is particularly appropriate when the dimensionality of the independent space i.e., number of input variables is high. For the reasons given above, Naïve Bayes can often outperform other more sophisticated classification methods (The Statistics Homepage, 2003). The Bayesian Theorem is given

by $P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$. Generally the problem is to find the hypothesis H that best explains the data D.

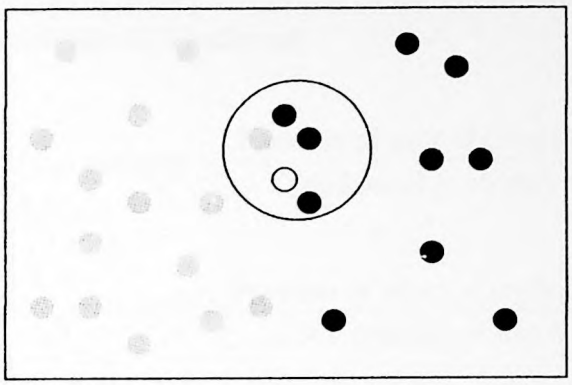

Figure 2.2: Classifying a New Object

In order for us to demonstrate the concept of the Naïve Bayes classification, consider the example shown in Figure 2.2. There are both grey and black objects. Our task is to classify the new object which is the white object (namely X). Since there are 15 grey objects and only 10 black objects in the figure, it is logical to believe that the new object is likely to have membership of grey rather than black. In the Bayesian analysis, this belief is known as the prior probability (The Statistics Homepage, 2003). So, the prior probabilities for grey circle and black object are:

$$\text{Prior Probability for grey object} = \frac{\text{Number of grey objects}}{\text{Total number of objects}} = \frac{15}{25}$$

$$\text{Prior Probability for black object} = \frac{\text{Number of black objects}}{\text{Total number of objects}} = \frac{10}{25}$$

We assume that the more grey (or black) object around X, the more likely that X belongs to that particular colour. So, in order for us to measure that, we draw a circle around X which encompasses a number of points irrespective of their colour labels. Then we calculate the number of objects in the circle belonging to each class label. From this we calculate the likelihood:

$$\text{Likelihood of X given grey} = \frac{\text{Number of grey objects in the circle}}{\text{Total number of grey objects}} = \frac{1}{15}$$

$$\text{Likelihood of X given black} = \frac{\text{Number of black objects in the circle}}{\text{Total number of black objects}} = \frac{3}{10}$$

Although the prior probability indicates that X may belong to grey but the likelihood indicates otherwise. In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule (The Statistics Homepage, 2003).

Posterior probability of X being grey

= Prior Probability for grey object × Likelihood of X given grey

$$= \frac{15}{25} \times \frac{1}{15} = \frac{1}{25}$$

Posterior probability of X being black

= Prior Probability for black object × Likelihood of X given black

$$= \frac{10}{25} \times \frac{3}{10} = \frac{3}{25}$$

Finally, we classify the X as black because it achieves the highest posterior probability.

From the above visualization, we can conclude that it is one classifier that is easy to comprehend. Naïve Bayes also easily handles missing values by simply omitting single attribute probabilities for each class. However, as the attributes of most of the datasets available are usually not all independent, this contradicts with Naïve Bayes' assumption and might affect the performance of this classifier.

2.3.4 K-Nearest Neighbours (KNN)

In this classification technique, a new variable with an unknown label is assigned the label of the variable in the training set which is nearest and similar. The nearest neighbour algorithm is extremely simple and is used in many applications. The similarity may be measured using distance measures which include Euclidean distance, Euclidean squared distance, Manhattan distance (also known as City-block distance or taxi-cab distance), and Chebychev distance.

While nearest neighbour refers to the nearest neighbour or 1 nearest neighbour, *k*-nearest neighbour or KNN refers to the *k*th nearest neighbour. Apart from that, KNN is a more robust method that classifies data points by looking at more than just the nearest neighbour. KNN is a memory-based method. That, in contrast to other statistical methods, requires no training. It functions on the intuitive idea that close objects are more likely to be in the same category. Thus, in KNN, predictions are based on a set of prototype examples that are used to predict new or unseen data based on the majority vote (The Statistics Homepage, 2003).

The KNN classifier in WEKA uses the Euclidean distance, D which is the distance measured between the sample with the gene values $a_1^{(1)}$, $a_2^{(1)}$, ..., $a_k^{(1)}$ (where k is the number of attributes) and one with values $a_1^{(2)}$, $a_2^{(2)}$, ..., $a_k^{(2)}$.

The formula is given by $D = \sqrt{(a_1^{(1)} - a_1^{(2)})^2 + (a_2^{(1)} - a_2^{(2)})^2 + ... + (a_k^{(1)} - a_k^{(2)})^2}$

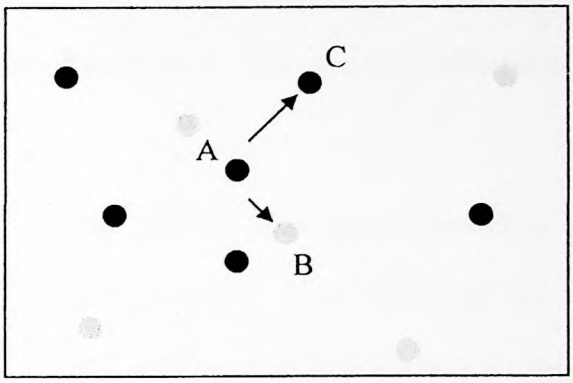

Figure 2.3: The Distance Between A-B and A-C

On the other hand, when nominal variable are present as shown in Figure 2.3, it is necessary to come up with a "distance" between different values of that variable. In this case, we have to calculate the distance between the black dots and the grey dots as seen in Figure 2.3. Usually a distance of 0 is assigned if the values are identical, otherwise the distance is 1. Thus, the distance between black and black is 0 and that between black and grey is 1.

If the value of k becomes very large, then the classification will become all the same – simply classify each attribute as the most numerous class. For this study, we will use k=4.

The KNN classifier is user-friendly and gives optimal results by numeric data. However, the weakness of this classifier is its large computing power

requirement, since the distance to all the objects in the dataset has to be calculated in order to do classification and the database also can be easily corrupted by noisy exemplars, which are the already-seen instances that are used for classification (Ye, 2004).

2.3.5 Support Vector Machine - Sequential Minimal Optimization (SMO)

Support vector machine (SVM) is a linear modeling that is used for classification and instance-based learning. It is based on the maximum margin hyperplane. SVM selects a small number of critical boundaries called support vector from each class and builds a linear discriminate function that separates them as widely as possible. Support vector is a set of points in the feature space that determines the boundary between objects of different class memberships. It transforms the instance space into a new space. With a nonlinear mapping, a straight line in the new space does not look straight in the original instance space. A linear model constructed in the new space can represent a nonlinear decision boundary in the original space (Witten and Frank, 2000).

If there is a two-class dataset whose classes are linearly separable; that is, if there is a hyperplane in instance space that classifies all training samples correctly then the maximum margin hyperplane is the one that gives the greatest separation between the classes.

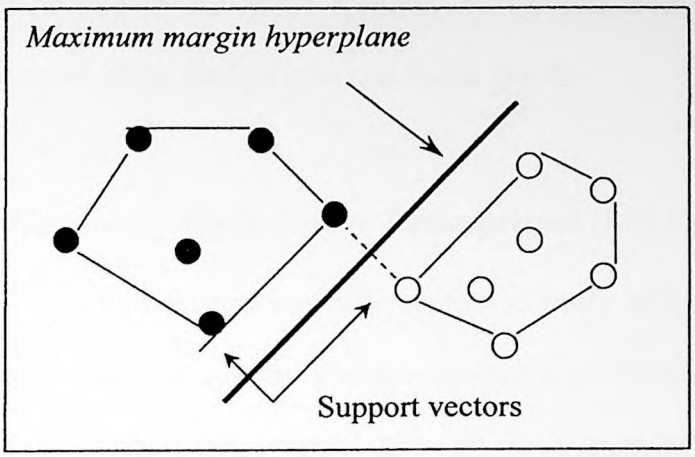

Figure 2.4: A Maximum Margin Hyperplane

In Figure 2.4, two classes are represented by open and filled circle. We connect each circle of the same class and two polygons are created. Since we assumed that the two classes are linearly separable, it cannot overlap each other. Among all hyperplanes that separate the classes, the maximum hyperplane is considered to be the one that is as far as possible from both the polygons that are built. The equation of the hyperplane separating the two classes can be written as $x = w_0 + w_1 a_1 + w_2 a_2$ with a_1 and a_2 as the variable values and w as weights to be learned.

The instance that is closest to the maximum margin hyperplane is the one with minimum distance and it is called the support vector. There is always at least one or more support vector for each class (Witten and Frank, 2000).

There are many methods to train SVM. One particularly simple method is Sequential Minimal Optimization (SMO) which is what we will be using in WEKA. Nevertheless, SMO is often slow to converge to a solution, particularly when the data

is not linearly separable in the space spanned by the nonlinear mapping. This situation increases with noisy data (Witten and Frank, 2000).

2.3.6 Neural Network - Multi Layer Perceptrons (MLP)

Neural network has been successfully applied in many areas. Indeed, neural network can be seen anywhere especially when it comes to problems like prediction, classification or control. Basically, neural network is so popular because of its powerful algorithm and the fact that it is easy to use. In addition, neural network is nonlinear and is also a very sophisticated modeling technique which is able to model an extremely complex function. However, the algorithm is also not as easily comprehensible as the others and is often called the black box.

The basic neural network consists of neurons. A neuron receives a number of inputs either from the original data or from the output of other neurons in the neural network and each of the input comes via a connection that has a strength or weight. Each neuron also has a single threshold value. The weighted sum of the inputs is formed, and the threshold is subtracted to compose the activation of the neuron. The activation signal is then passed through an activation function to produce the output of the neuron (The Statistics Homepage, 2003).

A simple network, as shown in Figure 2.5, has a feedforward structure: signals flow from inputs, forwards through any hidden units, eventually reaching the output units. A typical feedforward network has neurons arranged in a distinct layered topology. The input layer is not really neural: these units simply serve to introduce the values of the input variables. The hidden and output layer neurons are

each connected to all of the units in the preceding layer (The Statistics Homepage, 2003).

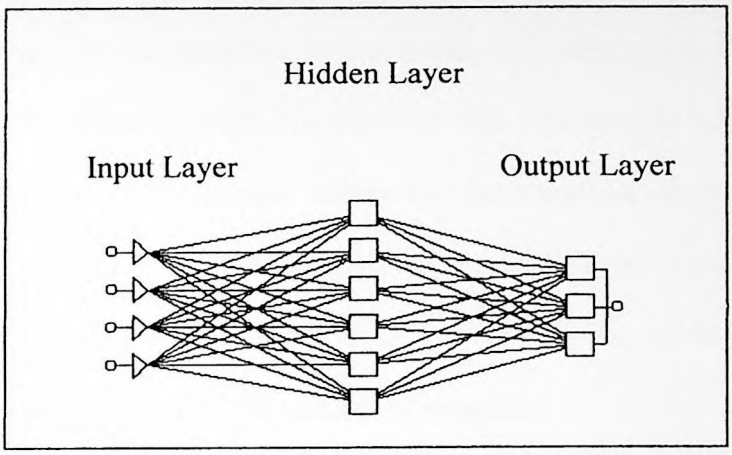

Figure 2.5: A Basic Artificial Model

When the network is used, the input variable values are placed in the input units, and then the hidden and output layer units are progressively executed. Each of them calculates its activation value by taking the weighted sum of the outputs of the units in the preceding layer, and subtracting the threshold. The activation value is passed through the activation function to produce the output of the neuron. When the entire network has been executed, the output of the output layer acts as the output of the entire network (The Statistics Homepage, 2003).

The neural network is trained using one of the supervised learning algorithms. The supervised learning networks are the Multi Layer Perceptron (MLP), the Cascade Correlation learning architecture, and Radial Basis Function networks (Michie et. al. 1994). However, the most popular network architecture in use is the MLP and will be used for this study. It also uses the concept and the algorithm that we discussed in the previous part. The number of input and output units are defined