Kuwait J. Sci. 46 (1) pp 24-32, 2019

Simple empirical models of classifying patients from microarray data

Alan Oxley
Engineering, Design and Information & Communications Technology (EDICT), Bahrain Polytechnic, PO Box 33349, Isa Town, Kingdom of Bahrain
**Corresponding author: alan.oxley@polytechnic.bh*

Abstract

There have been tremendous advances in bioinformatics in recent years. One of these is the use of microarrays for collecting Big Data. This paper reports on the work carried out by the author in devising models to classify patients by conducting microarray data analyses. The problem is to determine, for each patient, which class he/she belongs to. For example, one class may be "has the disease" while the other class is "does not have the disease." Membership of a class can aid in giving a patient a prognosis. Often only a small number of genes are significantly affected by the presence of a disease, so it is possible to classify a patient by looking at this small number of genes. Two models for classifying patients from gene expression microarray data were developed. One model involves an existing algorithm, while the other involves a new algorithm. The models involve some simple mathematical techniques (the two sample student's t-test, Diagonal Linear Discriminant Analysis) and a newly developed technique which shall be called Multiplicative Probabilistic Discriminant Analysis. Each model has been implemented as a computer program. The research restricted itself to one dataset. Prior to using the models, the raw data must be pre-processed.

Keywords: Cancer; disease classification; genes; gene filtering; microarray data .

1. Introduction

Computer-aided medicine comes in many forms (Thomas, 2016). One of these is the use of computer systems to trawl through huge datasets to find patterns. This data-driven approach to medicine is a type of artificial intelligence (AI). It has the potential to speed up and improve the accuracy of disease diagnosis.

1.1 Research rationale

This paper builds on the work of Einbeck *et al.* (2015). Oxley (2017) notes, "This describes the processing of two datasets. Both datasets contain numerical values that describe the genes of medical patients. One dataset is for breast cancer patients, while the other is for irritable bowel disease (IBD) patients. Einbeck *et al*. (2015) describe the development of a tool that learns from the patterns in the gene values of known patients. The finished tool can take a new patient's gene values and" determine which class of patients the new patient is most likely to belong to. "The cancer dataset is readily accessible. The author replicated the work done by Einbeck *et al.* (2015)." The IBD dataset is not very accessible, and so the author did not undertake any work with the IBD dataset. Oxley (2017) states "WhilstEinbeck *et al*. (2015) describe the processing of the dataset; the paper does not describe the pre-processing of the dataset. The author found that a considerable amount of work was involved in pre-processing the dataset."Einbeck *et al.* (2015) used a trio of basic statistical methods (t-test, correlation threshold, DLDA). Their accuracy rates of around 90% for breast cancer ER classification are comparable to the rates obtained in other studies. Similarly, they achieved accuracies of close to 70% in IBD studies, which is in line with the literature that deals with complex techniques.Much research has been undertaken into analyzing microarray gene expression data in order to classify patients. Before classification can commence, we must learn how the expression value patterns relate to each of the classes in question. For patients' gene expression values, different genes contribute different amounts to classification. A microarray dataset comprises a sample of subjects (patients). Consider a dataset, discussed later, of breast cancer patients whose lymph node status is negative. The dataset includes the expression values for the genes and which of two classes the patient belongs to—oestrogen-receptor positive (ER+) and oestrogen-receptor negative (ER-). We use this dataset to learn how patterns in the expression value data relate to the membership of the classes. As seen in Table 1, the expression values for the ER- class are generally higher than those for the ER+ class. Therefore, the expression values for the gene 201201_at can be used, along with other genes, to classify a patient whose expression values are known but whose class is not. In contrast, the expression values for gene 221706_s_at have a very similar spread, and so are of no use in predicting to which class a new patient (a future subject) belongs.

1.2 Research objectives

The objective was to develop two models for classifying patients from gene expression microarray data. One model was to involve an existing algorithm, whereas the other was to involve a new algorithm. A model for classifying patients from gene expres-

Table 1. Gene expression values for two genes. Left gene's values are generally different for ER+ and ER- groups; right gene's values are generally similar.

	Gene 201201_at		Gene 221706_s_at	
	ER-	**ER+**	**ER-**	**ER+**
Min	1571.3	2277.4	68.4	30.3
1 quart	5317.9	3838.2	229	220.5
Median	7462.1	4635.7	281.5	286.7
3 quart	10722	5523.2	350	356.6
Max	30962	15626	542.7	648.5

sion data consists of two main steps: a procedure for identifying a group of significant genes from "known" patients (gene selection) and a classifier that decides to which class each "new" patient belongs (classification).

We obviously needed to validate each model, i.e. see how good it is at classifying patients. A complication was that each model has two parameters whose optimum values were unknown. The research intended to run each model many times to see the effect of different parameter values on the accuracy with which 'new' patients are classified. One of these parameters is the number of significant genes to be considered. The other parameter is the value of the correlation coefficient. It is important to gain some insight into the optimum values to be used in a model. Only one dataset was used that included details of 286 patients who have suffered, or are suffering, from breast cancer. The number of genes that the microarray recorded as being present is 17,816. The ER status of each of these patients is known, i.e. either class ER+ or ER-. In addition, we know which patients have had a relapse and which patients have not. Thus, the "relapse" status of each patient is known, i.e. either "relapsed" or "not relapsed." We therefore can conduct two sets of independent experiments, one using the ER status data and one using the relapse status data.

1.3 Proposed solution

A microarray dataset is of a high dimension. Statistical methods exist to analyze the dataset for patterns and retain the number of dimensions, i.e. the number of genes. These methods can take a substantial amount of time to process, and, therefore, incur a high cost. Some methods reduce the number of genes before continuing with the processing. This latter strategy was selected. Both proposed models had the same gene selection step. This was to find the g most significant genes from the patient cohort. In this project, the approach used a two-sample t-test which identifies effects of the type shown in Table 1, as well as a correlation threshold to eliminate highly correlated genes. The gene selection step comprised three sub-steps:

a) Genes were first ranked according to their significance using the two-sample t-test. A gene whose values were generally quite different for one group of patients (e.g. ER+) than they were for the other group (e.g. ER-) had a high rank. Similarly, a gene whose values were generally the same for both groups of patients had a low rank (see Table 1). A list of significant genes could then be created.

b) The top ranking 100 or so genes were selected.

c) The expression values of some genes were closely related to the expression values of other genes. One explanation for this is that the genes have a related function. As only a small number of genes were selected in this study, then they should have been independent of one another. Therefore, starting from the second most significant gene, each gene in the list was compared in turn with those genes of higher rank. If the gene expression values of the pair of genes were closely correlated, the lower ranking gene was removed from the list. For closely correlated genes, each class of patients also had to be closely correlated. For example, for gene x and gene y, if the "relapse" class was closely correlated for both *x* and *y,* and the "not relapsed" class was closely correlated for both *x* and *y,* then genes x and y were closely correlated.

The classification step predicted the status of a new patient. For this step, one model used an existing algorithm—diagonal linear discriminant analysis (DLDA). Despite its name, DLDA is a simple statistical method. The other model used a new algorithm called the multiplicative probabilistic discriminant analysis (MPDA).

1.4 Description of the paper

This paper begins with a section on Background Work, which describes related theory, the technology used in the research, and related research. The Design section describes the creation of the two models. The Implementation section briefly discusses the use of large files. The discussion summarizes what results and their implications.

2. Background work

As the fields of Big Data and Bioinformatics are relatively new, this section provides some background information that will help with understanding the basis for the research.

2.1 Related theory

For an introduction of the subject of genetics see, for example, University of Utah (n.d.) and The Nemours Foundation (2017). The human body is made up of cells. Most cells have one nucleus. Within each cell nucleus are spaghetti-like structures called chromosomes. These come in matching pairs. There are 23 pairs of chromosomes, but not every living thing has 23 pairs.

Chromosomes have different lengths and patterns. Hundreds of thousands of genes are found on every chromosome. Each gene has a specific function, and genes too come in pairs. Genes are often associated with a person's traits. For example, if both of an individual's parents have green eyes, then the person might inherit green eyes from them. If a person's mother has one gene for brown hair and another for red hair, and the person has red hair, then it might have been inherited from the person's mother. Genes also account for traits in other animals, such as dogs.

Several years ago, scientists devised numerous ways to study genes: analyzing the proteins they encode, cloning them, making mutations in them, mapping them, and sequencing them. Scientists usually applied their studies to one or a small number of genes. Today, scientists can study all the genes in the human body at the same time. This is the science of genomics.

Almost every cell in the human body contains copies of the 25,000 to 35,000 genes. Each cell type has some genes turned "on" while others are turned "off". This list of what is on and what is off is referred to as the gene expression of a cell.

DNA microarray analysis is a highly active area of genetic research. A DNA microarray (aka 'DNA chip,' 'gene array,' and 'genome chip') is a device that. One such devise created by a biotechnology company is called GeneChip. A microarray is about the size of a packet of cigarettes. The device is created by a robot and contains about 20,000 different probes.

First, a patient's cells are collected. Then, through a complex laboratory process, the cells are analyzed. Thereafter, "it is possible to measure how each gene expresses itself" "Oxley (2017). The microarray measures how all of a patient's genes expresses themselves.

After microarray analysis, a single value associated with each gene is available. Consider a certain disease. If we look at the gene expression values of a group of people who have a disease and compare it with the values for a group of people who are well, then different patterns in the data emerge. A relatively small group of genes may be affected by the presence or absence of a disease.

Cancer is a disease where something has gone wrong with some of one's genes. We can perform an experiment to use a DNA microarray to measure the gene expression levels of cancer cells and compare them with the levels for healthy cells.

Mackintosh (2017) states, "Data saves lives… If health data is liberated and can be analyzed by the best minds and machines, warning signs could be spotted before they become full blown crises… Your data can save your friend, family and neighbor's lives and that needs communicating." Data has the ability to transform healthcare. Mackintosh (2017) argues that we need to espouse the sharing and probing of health data, and there is a cost to not doing so. We must ensure that the public is at ease with the idea of its data being shared.

Large databases which record the disease and treatment history for most patients in the developed world are available to the general public (Hall, 2016). The presence of big data allows us to use novel approaches, such as machine learning. DeepMind is one company that undertakes AI research into healthcare. Its parent company is Alphabet.

Consider the problem of trying to predict whether a recovering cancer patient is going to relapse. One way to do this is by consulting a dataset of patients. For each patient, the gene expression values are given together with the outcome of that patient, i.e. whether or not the patient relapsed. With this approach, we have developed a prediction model of patient outcome by learning from past patient data. An alternative approach might be to understand what part significant genes play in causing relapse, or not. A model can thus be built based on the resultant theory. If we had a data-based model and a theory-based model, then more accurate predictions could be made.

Empirical models (data-driven) and substantive models (theory-driven) require the values of some parameters to be set. The availability of big data encourages researchers to seek a data-based model by studying the patterns in the data. One author (Anonymous, 2016) has concerns about empirical models. A difficulty with the empirical model is that we do not know to what extent the model can be applied to different kinds of dataset. This is because we do not understand the underlying theory of how the model works. With an empirical model, a "black box" approach is used, and we rely on its performance at correctly classifying, e.g. whether a patient will/will not relapse, from known data.

There is also an ethical question with the empirical model. We use a model that we only superficially understand to tell a patient "You are likely to have a relapse," or the alternative. Anonymous (2016) argues that in patient prognosis, an automated decision-making system based on patterns of data, where little usage has been made of underlying theory, would seem to be inappropriate.

2.2 Related work

Oxley (2017) notes: "Einbeck *et al.* (2015) and Jackson *et al.* (2016) describe a project that … involves processing of the same cancer dataset that is being used in this paper (NCBI, 2016). It also involves processing an IBD dataset. Anyone can download the cancer dataset. However, downloading the IBD dataset is more complex and requires the use of a proprietary program. The papers by Einbeck *et al.* (2015) and Jackson *et al.* (2016) are understandable to the general reader who is unacquainted with microarray data analysis. The papers assume that the dataset has been pre-processed. The dataset contains gene expression values of several patients who have suffered/are suffering from breast cancer. For each patient, a variable specifies whether a patient has/has not relapsed. Furthermore, for each patient, a variable

specifies whether the patient is oestrogen-receptor positive (ER+) or oestrogen-receptor negative (ER-)." The papers described above make use of Discriminant Analysis.

The *t*-test is a null hypothesis testing procedure of the type known as a parametric test. In the context of this research, we have a sample of patients belonging to class A and a sample belonging to class B. (The two classes could be, for example, ER+ and ER-). For any gene, we look at the gene expression values of patients in A and compare them with the values of patients in B. We use the t-test to establish how likely it is that the difference between the samples is due to chance alone. In the gene selection step of this research we use the t-test to select those genes where the values in A differ so much from those in B that there is something of potential biological interest.

Discriminant analysis is a statistical technique that identifies the properties of two or more groups of objects with the aim of distinguishing between the groups. Having done this, when presented with a new object, a decision can be made as to the group to which the object should belong.

Consider DLDA as used in the context discussed here. For each gene, it is assumed that the gene expression values for all "known" patients in one class follow a normal distribution. Similarly, it is assumed that those in the other class follow a normal distribution. A description of DLDA now follows. DLDA involves:

Consider several patients whose statuses are known. Each is either in class A or B. Assume that we also know the gene expression values for each patient. We can represent this information as a table in which each row corresponds to a specific gene and each column corresponds to a specific patient. Each cell of the table contains a number. Cell (i, j) holds the expression value of gene i of patient j. Assume that all patients in class A occupy the left-hand side of the table, and all class B patients occupy the right-hand side. Note that we only use a small number of genes (g genes), those which our analysis indicates are responsible for classifying a patient's status.

1. Gene i, which is one of g significant genes, is considering. We have two normal distributions of the gene expression values: one for patients belonging to class A and one for those belonging to class B, as shown in Figure 1.

2. Each distribution is moved independently so that its mean is zero, as shown in Figure 2. Consider a value x from the original distribution. For New Group A, its *x*'-value is $(x - \mu_A)$, where μ_A is the mean of the original class A distribution.

3. The standard deviation is calculated for the combined distributions, $\sigma_{All,i}$.

4. The *x*-values of the original distribution is taken and a normalisation process is performed to give *x*"-values. Each x-value (the original gene expression value) is divided by $\sigma_{All,i}$, as shown in Figure 3.

5. Steps 1 to 4 are repeated for all g significant genes.

6. Calculate whether a new patient is a member of class A or a member of class B. We take the normalized gene expression values for the new patient $(x_1", x_2", ..., x_g")$ and calculate d_A^2 and d_B^2 as follows:

$$d_A{}^2 = \sum_{i=1}^{g} \left(x_i" - \mu_A"\right)^2 \; ; \; d_B{}^2 = \sum_{i=1}^{g} \left(x_i" - \mu_B"\right)^2$$

where $\mu_A"$ and $\mu_B"$ are the means of the normalised class A distribution and normalised class B distribution, respectively. If d_A^2 is smaller than d_B^2, then the patient belongs to class A. Otherwise the patient belongs to class B.

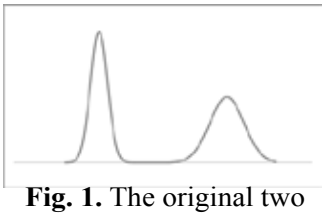


Fig. 1. The original two normal distributions.

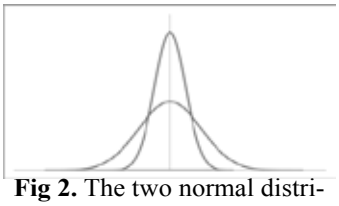


Fig 2. The two normal distributions each with their means moved to zero.

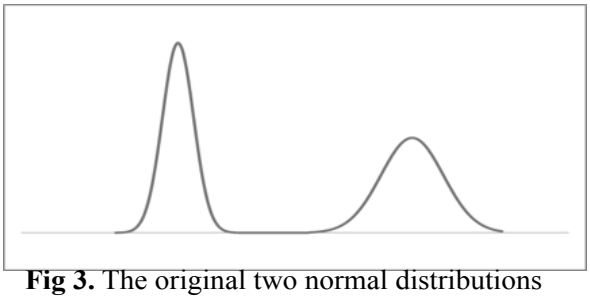


Fig 3. The original two normal distributions after being normalized.

2.3 Technology

Oxley (2017) describes the dataset: "The dataset prior to pre-processing comprises a single file of 46 Mbytes. Pre-processing involves three main steps. The dataset contains much extraneous data and information, so the first step is to get rid of the surplus information, leaving only a table where each column represents a patient and each row represents a gene. In addition, the body of the table comprises gene expression values—one value per table cell. The second step is to remove some of the genes, i.e. some of the rows of the table. This is done because when the microarray was used, for certain genes it could not detect anything for almost all the patients. However, in another database (244 Mbytes), information on which genes these are is available.

The third step involves creating two files. Initially, each file data is in one table, as just described. Let us call these files the "relapse table" and the "ER table." Consider the "relapse table." There is information on a webpage showing the relapse status of each patient, i.e. whether the patient has had/ has not had a relapse. Using this information, we sort the columns of the relapse

table representing the patients, according to their relapse status. For the ER table, there is information on the same webpage showing the ER status of each patient, i.e. whether a patient is ER+ or ER-. Using this information, we sort the columns o according to the ER status. After pre-processing, two files each of size 29 Mbytes have been created.

Processing involves writing computer programs to study the table and developing tools that give a prognosis for a new patient. Each program can be executed independently with both the relapse table and the ER table. The required technology for the research is shown in Table 2.

Table 2. Resources and technology used in the research (Oxley, 2017).

Stage	Purpose	Resource / Technology
	The database containing the dataset	Website
	The dataset	A 46-Mbyte file that the author has downloaded from the website and installed on his computer's hard disk.
Pre-processing	Step 1: Tidying up	Spreadsheet program, e.g. Microsoft Excel
	Step 2: Removing some genes	A 244-Mbyte file that the author has downloaded from the website and installed on his computer's hard disk. Spreadsheet program. Programming language, e.g. Octave
	Step 3: Creating relapse table and ER table files	Website Spreadsheet program
Processing	Processing the relapse table file	A 29-Mbyte file that the author created; it is the result of pre-processing the dataset. Programming language
	Processing the ER table file	Another 29-Mbyte file that the author created; it is the result of pre-processing the dataset. Programming language

All the processing work done in this research was achieved by writing Octave programs. The diagrams in this paper involved taking program outputs and plotting the data using a spreadsheet program.

3. Solution design

Wang *et al*. (2005) state that the data is in the NCBI/Genbank GEO database (series entry GSE 2034). The URL is https://www.ncbi.nlm.nih.gov/genbank/. At the completion of the whole of the pre-processing tasks, we end up with the relapse and ER table files. The relapse table and ER table have over 17,000 rows, each corresponding to a specific gene.

Table 3. Representation of the dataset. ('g.e.v.' means 'gene expression value.')

	Patient 1	**Patient 2**	**…**
Gene 1	g.e.v.	g.e.v.	
Gene 2	g.e.v.	g.e.v.	
…			

3.1 Gene selection

In order to come up with a list of significant genes, the genes are first ranked. There are many ways in which ranking can be done. Many of these involve ranking according to the value of a statistic. The method chosen is the two-sample t-test; the higher the test statistic the higher the significance of the gene. Only a relatively small number of table rows account for the status of a patient. Furthermore, genes that are closely correlated with a higher-ranking gene are removed. Thus only g genes are selected prior to subsequent processing. For a pair of genes to be correlated, the gene expression values for one class must be closely correlated and the gene expression values for the other class must be closely correlated. The optimum value of the correlation coefficient is not known. Similarly, the optimum number of significant genes to be used is not known. In order to ascertain the optimum values for these parameters, the models were executed several times for different correlation coefficients and different numbers of significant genes. The correlation coefficient varied from 0.6 to 1 in steps of 0.1. The number of significant genes varied from 5 to 60, in steps of 5.

3.2 Classification

After a small number of genes have been selected, a classifier needs to be used. There will be expression values at its input, for selected genes, for the "known" and "new" patients. The class that

the "new" patient belongs to will be at its output. The DLDA model uses the existing algorithm for classification. The MPDA model uses the new algorithm for classification.

3.3 Algorithm

This algorithm processes the same table as is described in Section 2.2, at the start of the DLDA algorithm description. MPDA involves:

1. Consider gene *i*, which is one of g significant genes. We have two normal distributions of the gene expression values, one for patients belonging to class A and one for class B, as shown in Figure 1.

2. Normalize the gene expression values as with DLDA (see steps 2 to 4 of the process described earlier).

3. Divide the axis into equally sized intervals.

4. Locate the interval which contains the new patient's normalized expression value, *xi"*. Calculate the area of a strip under the normalized normal distribution of class A for this interval, as shown in Figure 4. Denote this area by $area_i$.

5. Repeat steps 1 to 4 for all g significant genes.

6. Calculate the product of the areas

$$P_A = \prod_{i=1}^{g} area_i .$$

7. Repeat steps 1 to 6 but now for the normalized normal distribution of class B.

This gives $P_B = \prod_{i=1}^{g} area_i$.

8. Calculate whether a new patient is a member of class A or a member of class B. PA and PB relate to the probabilities that the new patient is a member of classes A and B, respectively. If PA > PB then the new patient is a member of class A, otherwise the patient is a member of class B.

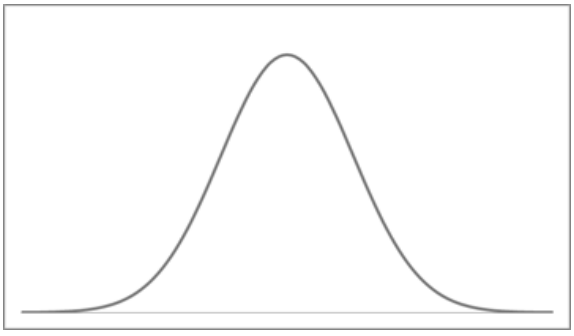


Fig. 4. Identifying the area of a strip under the normalized A distribution.

4. Testing

4.1 Participants

The dataset only contains the statuses of patients whose ER and relapse statuses are known. In order to get "new" patients, the statuses of some of the patients were assumed to be "unknown". Approximately a quarter of the patients were randomly re-categorized as "new." Let us refer to the two groups of patients as "known" patients and "new" patients, or training set and the test set, respectively. The list of genes was formulated only using the statuses of the "known" patients.

4.2 Test plan

Having selected a list of genes and a method of classification, the next step is to test the models to see how accurately each classifies 'new' patients. In other words,. the models must be validated. To summarize, the following tests were carried out:

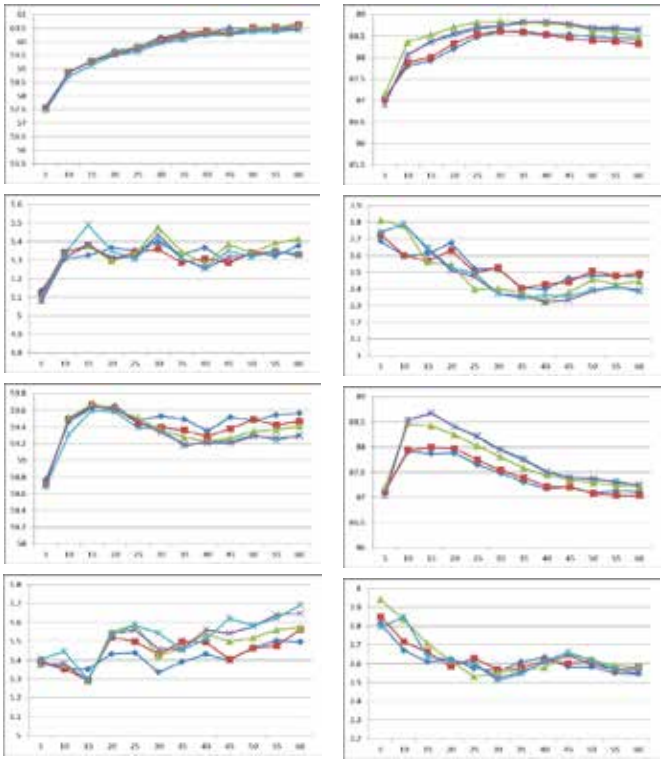
- Model using DLDA
  - o ER status data: 60 combinations (12 different gene list lengths; 5 different correlation coefficients). Each combination repeated 1,000 times, each time the "new" patients were randomly chosen from scratch. For each of the 1,000 runs, the proportion of "new" patients whose statuses were classified correctly was recorded. The results for 1,000 runs of each combination were then averaged, and the standard error was calculated.
  - o Relapse status data: 60 combinations. For each combination, 1,000 runs, the results were averaged, and the standard error was calculated.
- Model using MPDA
  - o ER status data: 60 combinations. For each combination, 1,000 runs, the results were averaged and the standard error was calculated.
  - o Relapse status data: 60 combinations. For each combination, 1,000 runs, the results were averaged and the standard error was calculated.

4.3 Results

The DLDA-based model was run 1,000 times for each of the relapse status data and the ER status data. Similarly, the MPDA-based model was run 1,000 times for each of the relapse status data and the ER status data. Figure 5 shows the results of varying the number of genes selected and the correlation coefficient value, for both models. Figure 6 shows all results of the mean percentage of "new" patients whose statuses has been correctly predicted.

The number of occurrences of each gene was tallied. This was done by considering the the highest-ranking 60 genes for each of the five correlation coefficient values (i.e. the 60 best genes after closely correlated genes had been removed). The result was a total of 300. This was repeated for each of the 1,000 times that the program was executed, equaling a cumulative total of 300,000. For the relapse table, the highest tallies occurred for the following genes (from highest tally to lowest):

209380_s_at; 202324_s_at; 218252_at; 219312_s_at; 222077_s_at; 206188_at; 212149_at; 212898_at; 202824_s_at; 214853_s_at; 218478_s_at; 212900_at; 209831_x_at; 218701_at; 219215_s_at; 201076_at; 32088_at; 213391_at; 211004_s_at; 201368_at.



Symbol	♦	■	▲	×	✱
Correlation coefficient	0.6	0.7	0.8	0.9	1.0

Fig. 5. Mean percentage correctly predicted (left column) and standard error (right column) for relapse status (DLDA: top row; MPDA: 2nd row) and ER status (DLDA: 3rd row; MPDA: bottom row).

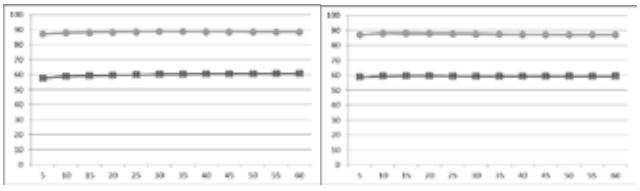


Fig. 6. Mean percentage correctly predicted for DLDA (left) and MPDA (right). Lower graphs are for relapse status; upper graphs are for ER status.

For the ER table, the highest tallies occurred for the following genes:

209602_s_at; 205225_at; 200600_at; 209791_at; 206074_s_at; 200827_at; 218211_s_at; 202088_at; 212956_at; 212195_at; 204862_s_at; 200711_s_at; 203256_at; 200670_at; 204667_at; 201231_s_at; 219497_s_at; 216237_s_at; 209191_at; 201579_at.

5. Discussion

5.1 Summary of achieved objectives

The objectives of this research have been achieved. The difficulty with writing this paper has been not to overwhelm the reader with the minutiae of how the dataset is pre-processed and processed. The work is current as it emanates from a recent publication Einbeck *et al.* (2015).

5.2 Future work

Future work should include upgrades and modifications to the program. When formulating a list of genes, prior to classification, a high-ranking gene will not be added to the list if it is closely correlated with a high-ranking gene already in the list. Rather than using this approach one could, at the outset, find groups of closely correlated genes and then, for each group, decide which of the genes is to be selected—it may be the gene that best represents the group. Irritable bowel syndrome (IBS) is a disorder affecting a large number of people. It would be useful to run both models with IBS data.

5.3 Conclusion

One microarray dataset has been studied. To classify the relapse and ER statuses of "new" patients, two models have been used. One used the *t*-test, a correlation coefficient, and the DLDA algorithm. The other used the t-test, a correlation coefficient, and the MPDA algorithm. For the MPDA-based model, the proportion of "new" patients whose statuses were correctly predicted is comparable to the DLDA-based model.

There are two parameters in each model that are used in gene selection: the number of genes to be selected (g) and the value of the correlation coefficient. Both parameters have been varied in order to see their effects on the results. The smaller the value of g, the more important it is to remove correlated genes, i.e. redundancy is costlier with a shorter list. It is possible that the magnitude of the correlation coefficient is not important when g is large. It may be that the optimum value of the correlation coefficient is dependent on the dataset, and so it must be estimated for each dataset. The research shows that the DLDA-based model and the MPDA-based model, even though they are relatively simple, can be used with a small number of genes. They produce prediction rates for the relapse and ER statuses of breast cancer patients that are comparable to more complex methods found in previous studies. The prediction rate for ER status is particularly high.

References

Anonymous. (2016). A doctor writes deep learning but

no deeper understanding. Mathematics Today, 52(6): 266.

Einbeck, J., Jackson, S. & Kasim, A. (2015). A summer with genes: simple disease classification from microarray data. Mathematics Today, 51(4): 186-188.

Hall, M. (2016). Conquering cancer. ITNOW, Sept: 40-41.

Jackson, S.E., Einbeck, J., Kasim, A. & Talloen, W. (2016). the correlation threshold as a strategy for gene filtering, with application to irritable bowel syndrome and breast cancer microarray data. Reinvention, 9(2).

Mackintosh, M. (2017). Data is the future of healthcare. ITNOW, March: 40-41. NCBI. Series GSE2034. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034 (3/10/16) Last accessed on 13/6/17.

Oxley A. (2017). designing a course on processing big data. 2017 e-Learnit Conference, Bahrain Polytechnic, Bahrain. The Nemours Foundation. (2014). What is a Gene? Retrieved from http://kidshealth.org/en/kids/what-is-gene.html (2017) Last accessed on 3/10/17.

Thomas, D.J. (2016). Computer-aided medicine revolution. ITNOW, Dec: 40-41. University of Utah (n.d.). Genetic science learning centre. Learn: Genetics. Retrieved from http://learn.genetics.utah.edu/ Last accessed on 3/10/17.

Wang, Y., Klijn, J.G., Zhang, Y, Sieuwerts, A.M., Look, M.P. et al. (2005). Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet, 365(9460): 671-679.

Submitted: 10-10-2017
Revised: 24-12-2017
Accepted: 04-03-2018

نماذج تجريبية بسيطة لتصنيف المرضى عن طريق تحليل بيانات المصفوفات الدقيقة

ألان أوكسلي

الهندسة والتصميم وتكنولوجيا المعلومات والاتصالات (EDICT)، بوليتكنك البحرين
ص.ب. 33349، مدينة عيسى، مملكة البحرين

alan.oxley@polytechnic.bh

الملخص

في السنوات الأخيرة، كان هناك تقدم هائل في المعلوماتية الحيوية، مثل استخدام المصفوفات الدقيقة لجمع البيانات الضخمة. ويقدم هذا البحث تقريراً عن الأعمال التي تم تنفيذها لابتكار نماذج لتصنيف المرضى من خلال تحليل بيانات المصفوفات الدقيقة. تكمن المشكلة في تحديد الفئة التي ينتمي إليها كل مريض. فعلى سبيل المثال، قد تنقسم الفئات إلى "مصاب بالمرض" و "غير مصاب بالمرض". وغالباً ما يتأثر عدد قليل من الجينات بشكل كبير بالمرض وبالتالي يمكن تصنيف المريض بالنظر إلى هذه الجينات. تم تطوير نموذجين لتصنيف المرضى من خلال تحليل بيانات المصفوفات الدقيقة الجينية. ويتضمن أحد هذه النماذج خوارزمية موجودة بالفعل بينما يتضمن النموذج الآخر خوارزمية جديدة. وتتضمن النماذج بعض التقنيات الرياضية البسيطة، مثل: اختبار "تي" للطالب ذو عينتين، وتحليل التباين الخطي القُطري وتقنية مُطورة حديثاً يُطلق عليها تحليل التباين الاحتمالي الضربي. تم تنفيذ كل نموذج كبرنامج حاسوب. واقتصر البحث على مجموعة بيانات واحدة. ويجب معالجة البيانات الأولية مسبقاً قبل استخدام هذه النماذج.