BS9001: Research Experience Report

Title: Meta-analysis of breast cancer using data science techniques

Justin Chia Cheng Ze

U1740466J

**Abstract**

Conducting a Meta-Analysis is crucial in determining the differences between studies on breast cancer. The aim of this study is to extract, clean, aggregate and visualize data on a single dataset using multiple data science techniques. The statistical programming language used in this study is R. All genes expressed in breast cancer patients were accessed via the Gene Expression Omnibus (GEO) database repository. The platform obtained from is from Affymetrix Human Genome U133 Plus 2.0 Array, GPL570. 20 gene expression profiles were analysed, with 4 patients with normal breasts and 16 patients with Invasive Ductal Carcinoma breasts from GSE22544 series. The results show high precision, recall, and F1-score when a single sampling was tested against a bootstrapped sample. The gene expression data was consistent and robust across all breast cancer specimens analysed.

**Introduction (1 to 3 paragraphs)**

Meta-analysis is essentially a statistical analysis that condenses the results of several scientific studies.1 Benefits of performing a meta-analysis is the combination of data, resulting in a higher statistical power and more robust point estimates than from individual studies.1 The steps in a meta-analysis include formulation of the research question, literature research, selection of studies, decision of which dependent variables are allowed, selection of a meta-analysis model, and examination of heterogeneity between studies.3 In this study, the stage of progress is done until the selection of studies.

Breast cancer is the development of cancerous tissue from breast tissue.4 Among breast cancers, cancers of the ducts are known as ductal carcinomas, and cancers of the lobules are known as lobular carcinomas.2 In this study, the type of breast cancer studied is known as Invasive Ductal Carcinoma or Infiltrating Ductal Carcinoma (IDC). IDC is the most common type of breast cancer, causing up to 80% of breast cancer cases.2

Cancer is a leading cause of death, with breast cancer being the most common cancer among women, totalling two million cases worldwide.2 Therefore, it is important to understand the genetic profiling of cancer among various studies to pinpoint the various significant genes that are highly associated with IDC. In this study, the aim is to obtain genetic expression data from genetic databases, to clean and manipulate it for use in the future meta-analysis.

**Methods**

**Selection of dataset/series from the GEO database**

A search was done on the GEO online database for suitable datasets/series from studies on invasive ductal carcinoma. Any suitable datasets/series were selected, and its corresponding accession number was recorded.

**Extraction of data from database**

The gene expression profile is sourced from the GEO database via the GEO accession number and manipulated using R. The working directory is first set. The GEO GSE file is obtained and assigned to a variable. The pheno data and feature data of the eSet is assigned to their respective vectors. Expression is then obtained from the feature data via the exprs() function. The gene symbols were then sub-setted from feature data and assigned to a variable as a vector. The expression data is then combined with the gene symbol vector to form a combined data frame. Row names were replaced with gene symbols. A class factor was created for further use in the bootstrap method. A gene symbols vector was created. All data frames, factors and vectors were written into files using the write.csv() command

**Compression of Gene Expression data**

The combined data frame and gene symbols vector were first read into the R script using the read.csv() command. Using the unique() function, the gene symbol vector was removed of any duplicate genes. An empty list was created using the list() function. A simple function was written to determine the mean expression level for genes which have multiple rows. The lapply() function was used on the list of gene symbols and expression data, with the beforementioned user-defined function. The list produced from the output was then unlisted using the sapply() function, and converted into a data frame. The data frame was then written into a file for further used.

**Bootstrap method**

The bootstrap method is done on the statistical programming language R in RStudio. The working directory is first set at the start of the R script. The data frame containing the gene expression data was then read into the R script. The data frame was then reorganised to change the “normal” class columns to the first 4 columns and the “IDC” class columns to the last 16 columns. 2 methods were employed to obtain a binary matrix of significant genes for 1000 resamplings (bootstraps). Method 1: Using nested ‘for’ loops to produce binary matrix. Method 2: Using “genefilter” package’s “rowttests” function. Method 2 was selected for its quicker running speed.

The resulting output stored in a list was then converted into a matrix by both the unlist() function and the matrix() function. The frequency of significance of each gene out of 1000 resamplings was computed using a ‘for’ loop. The “caret” package was then installed. A confusion matrix was then produced using the confusionMatrix() function from the “caret” package. The precision, recall and F-score were subsequently calculated using the output of the confusionMatrix() function. A visualisation of the confusion matrix was then created by producing a simple plot with a user-defined function. A distance matrix was then computed using a Jaccard’s coefficient user-defined function, and a nested ‘for’ loop. The data frames and matrices were then written into files for further use.

A screenshot of a cell phone

Description automatically generated**Results**

*Figure 1: Density plot of* sum\_vectvisualises the distribution of the number significant samplings for each gene.

With reference to figure 1, genes expressed in breast cancer are expressed at a highest density of around 50 - 100 samplings out of 1000. The frequency of gene significance decreases as the number of significant samplings increases.

**A screenshot of a cell phone

Description automatically generated**

*Diagram 1: The Confusion Matrix* and important metrics

Diagram 1 visualises the confusion matrix generated from comparison of a single sampling compared to that the summary of 1000 samplings. The precision, recall and F-score are high with values of 0.985, 0.945 and 0.965 respectively

**Discussion**

The results above show that there was high precision, high recall and a high F1-score. The analysis of the data may indicate that the bootstrap technique produces reproducible results as evidenced by the high recall. The high recall value of 0.952 reveals that there is a high number of relevant items selected. The high precision value of 0.926 reveals that there is a high number of selected items that are relevant. The F1-score is also the weighted average of the Precision and Recall. The high value of the F1-score, 0.0939, indicates good performance when comparing between ground truth and the 1000-times-bootstrapped sample. This may in turn, mean that the gene expression data is consistent and robust across all breast cancer specimens analysed. Another possible interpretation of the high precision and recall could be that the threshold of >180 significant gene expression observations for the bootstrapped sample was suitable for determining significantly expressed genes in breast cancer.

**References**

1. Fagard RH, Staessen JA, Thijs L. Advantages and disadvantages of the meta-analysis approach. J Hypertens Suppl. 1996 Sep;14(2):S9-12; discussion S13.
2. Sharma GN, Dave R, Sanadya J, Sharma P, and Sharma KK. Various Types and Management Of Breast Cancer: An Overview. J Adv Pharm Technol Res. 2010 Apr-Jun; 1(2): 109–126.
3. Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from <http://handbook.cochrane.org>.
4. Key TJ, Verkasalo PK, Banks E. Epidemiology of breast cancer. 2001. Lancet Oncol 2(3):133-140.