Indonesian Journal of Electrical Engineering and Computer Science
Vol. 11, No. 3, September 2018, pp. 1223~1227
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v11.i3.pp1223-1227

An Effective Pre-Processing Phase for Gene Expression Classification

Choon Sen Seah[1], Shahreen Kasim[2], Mohd Farhan Md Fudzee[3], Mohd Saberi Mohamad[4], Rd Rohmat Saedudin[5], Rohayanti Hassan[6], Mohd Arfian Ismail[7], Rodziah Atan[8]
[1,2,3]Soft Computing and Data Mining Centre, Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Malaysia
[4]Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Karung Berkunci 01, 16300, Bachok, Kelantan, Malaysia
[5]School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia
[6]Laboratory of Biodiversity and Bioinformatics, Universiti Teknologi Malaysia, 81300 Skudai, Johor, Malaysia
[7]Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Pahang, Malaysia
[8]Department of Software Engineering & Information System, Faculty of Computer Science and Information Technology, University Putra Malaysia (UPM), 43400 Selangor, Serdang, Malaysia

Article Info

Article history:

Received Apr 5, 2018
Revised Jun 6, 2018
Accepted Jun 20, 2018

Keywords:

Bioconductor
Data pre-processing
Gene expression dataset
Significant directed random walk

ABSTRACT

A raw dataset prepared by researchers comes with a lot of information. Whether the information is usefull or not, completely depends on the requirement and purposes. In machine learning, data pre-processing is the very initial stage. It is a must to make sure the dataset is totally suitable for the requirement. In significant directed random walk (sDRW), there are three steps in data pre-processing stage. First, we remove unwanted attributes, missing value and proper arrangement, followed by normalization of the expression value and lastly, filtering method is applied. The first two steps are completed by Bioconductor package while the last step is works in sDRW.

Corresponding Author:

Choon Sen Seah,
Soft Computing and Data Mining Centre,
Faculty of Computer Sciences and Information Technology,
Universiti Tun Hussein Onn Malaysia.
Email: seanseah0702@gmail.com

1. INTRODUCTION

Microarray technology is a branch of biology technology which aims to study the expression of genes from the cell [1]. It places the gene sequences on a glass slide called gene chip. The gene chip is designed to display the sequences of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Complementary base pairing between the sample cell and gene sequences on the chip produces different colours based on the expression level of the gene. The introduction of microarray technology allows researchers to analyse thousands of gene expression profiles simultaneously [2-5]. The datasets produced by microarray technology is known as gene expression dataset [2]. Much biomedical research, especially cancerous research, has been increased. However, the properties of large dimension would affect the result of research as well. Since the microarray dataset is large dimension, classifying and computing the algorithms becomes more complex to study the gene expression characteristics [6].

Besides that, microarray datasets have many improper attributes and missing values might occur after the first collection of dataset. The accuracy of the classification algorithm would be affected.

Hence, data pre-processing is one of the mandatory processes to undergo before the dataset can be applied into other mainstream research algorithms [7].

In the next section, we would like to introduce the used of gene expression dataset and its information, followed by the method to pre-process the dataset. While in section 3, the outcome after pre-processing the data will be demonstrated and a comparison will be made to showcase the difference before and after pre-processing of dataset. Lastly, we would like to conclude with the outcome before the ending of this research paper.

2. RESEARCH MATERIAL & METHODOLOGY

In this section, the material and methodology applied in the study will be explained. Gene expression dataset is applied as input dataset for the data pre-processing purpose. In this study, Bioconductor and significant directed random walk (sDRW) will play the role to pre-process the data before it has been used to predict and classified the genes.

2.1. Microarray Data

Gene expression dataset is the dataset produced by microarray technology. It can be accessed from National Center for Biotechnology Information (NCBI). In this research, GSE10072 [8] is downloaded in raw file. The platform to prepare this Affymetrix microarray gene expression dataset is GPL96. The samples identification (ID) of lung cancer dataset are between GSM254625 to GSM254731. GSE10072 is one of the lung cancer type samples set. It has 107 samples, of which 58 are cancerous samples and 49 are normal samples. In overall, GSE10072 has 13788 genes.

2.2. Methodology

A raw file of gene expression data is saturated with an abundance of information extracted from the cell. This raw file needs to be processed in order for the right attributes to be extracted for the next research study. R programming language is chosen and hence packages that are build up by R programming language will be used to pre-process the dataset. In our study, the Bioconductor package is downloaded and imported for the purposed of data pre-processing [9]. The Bioconductor will analyse the expression value and further arrange the dataset using normalization which narrows the range of data to be studied. In this study, the dataset will undergo 3 pre-processing stages before being applied into the real classification algorithms such as genetic algorithm [10], pathway based cancer classification [11], and significant directed random walk (sDRW) [12]. Figure 1 illustrate the phases in data pre-processing stage.

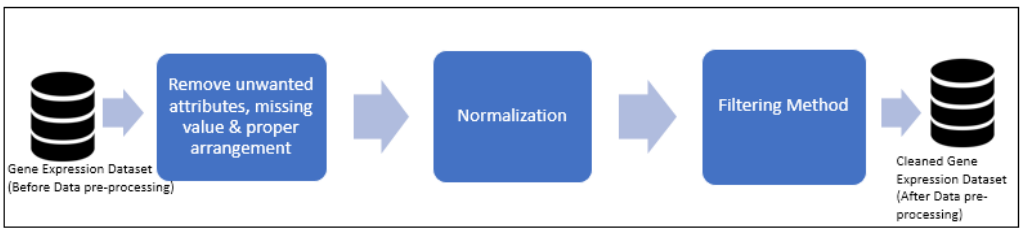

Figure 1. Data pre-processing of gene expression dataset

The first 2 steps will run under Bioconductor while the last step will run under sDRW. First, unwanted attributes, missing value and proper arrangements of dataset will be applied in order to clean the data. Figure 2 shows the details of the first step of data pre-processing. Unwanted attributes and the samples that have missing values will be removed. Rearrangement of data according to the requirement of format will be run through before proceeding to the next phases. Other than expression value, there is other information (attributes) such as patient biological information and dataset information which included in the gene expression dataset. All of this information is not going to apply in sDRW for cancer classification purposes [13]. Hence, these attributes are considered as unwanted attributes and will be remove from the dataset. Only wanted attributes will be kept for the cancer classification purposes [14], [15]. Example of wanted attributes are expression values, the position of genes, the means of gene's weight, and so on.

Attributes that have missing values will be prohibited because without the expression value, the deoxyribonucleic acid (DNA) sequences could not determine the actual expression value and the result will be affected if predicted value is applied to it. Hence, attributes that contain missing values will be

eliminated. Besides that, proper arrangement of datasets will be applied. DNA sequences are after each other and the "X" and "Y" value will determine the placing of the genes. The "X" and "Y" axis will determine the position of the genes in the sequences of DNA. This is important for the next process in sDRW because this could help in determining the next gene and hence, further referencing could be made by comparing with other reference data such as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and protein-protein interaction (PPI) sequences. Other than this, an additional of 3 values (mean, standard deviation and npixels) are used as additional attributes in cancer prediction and classification process. The mean is defined as the average of the sum of the weight of the gene. Standard deviation is the parameter which is used to quantify the amount of variation in the gene's weight.The npixels are the linear dimension of the genes in pixels.

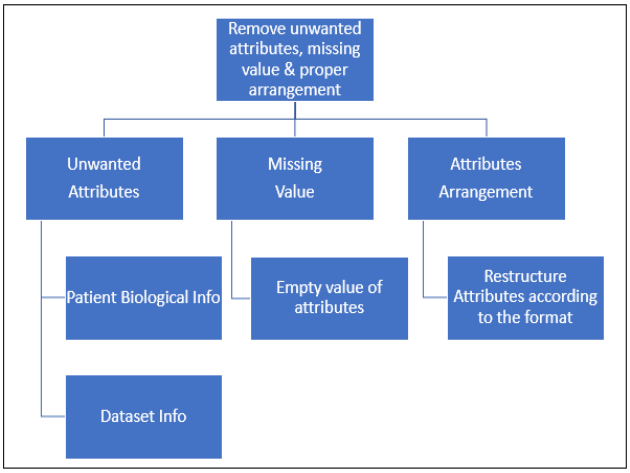

Figure 2. Step 1, remove unwanted attributes, missing value & proper arrangement

Second, normalization is applied to degrade the big value and causes the weight of the genes fall into the range between 0 – 10 [16]. During the normalization phase, gene's weight, means, and standard deviation is used to calculate the normalized value. For the normalized values that are greater than 10, it will be removed as we just want to keep the digit within 0 to 10. This is to remove the insignificant genes. Figure 3 illustrate the steps in normalization.

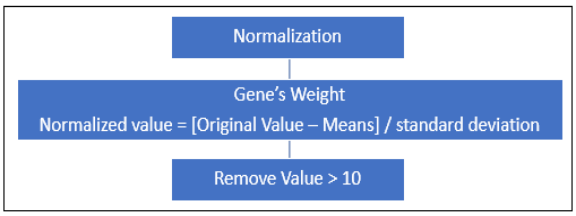

Figure 3. Step 2, normalization

Lastly, filtering method [17] is applied to select those genes that have p-value less than 0.05. This is because p value will determine the significant towards cancer mutation. Figure 4 shows the steps in filtering method. This step will be take places in sDRW.

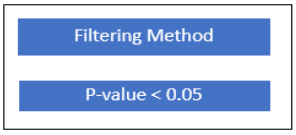

Figure 4. Step 3, filtering method

After having gone through the data pre-processing stage, the cleaned dataset is now ready to be applied in the evaluation algorithm and further implemented in classifier as well. Data pre-processing not only clears up the dataset to be ready for the implementation purpose but also allow the researchers to select the right attributes that would be the key influences for the study. In sDRW [12], Seah believes the weight of genes plays an important role in affecting the tumour formation and hence, during data pre-processing stage, he was focusing on those attribute that are related to gene's weight.

3. RESULTS AND ANALYSIS

In order to showcase the data pre-processing by Bioconductor and sDRW, lung cancer dataset, GSE10072 is used as an example. Dataset will be processed step by step according to the sequences arrangement in methodology. Originally, a raw file contains much information about the dataset which includes the unwanted information for the cancer classification process. But after data pre-processing, only wanted attributes are kept and being a further process in sDRW for the cancerous gene prediction and cancer classification purposes. Figure 5 shows part of the visualization of dataset in raw type. While Figure 6 illustrates the outcome of dataset after data pre-processing.

10	GridCornerUR=4492 212
11	GridCornerLR=4509 4482
12	GridCornerLL=248 4501
13	Axis-invertX=0
14	AxisInvertY=0
15	swapXY=0
16	DatHeader=[0..46115] NP2002061801_01:CLS=473...
17	Algorithm=Percentile
18	AlgorithmParameters=Percentile:75;CellMargin:2;Outli...
19	[INTENSITY]
20	NumberCells=506944
21	CellHeader=X Y MEAN STDV NPIXELS
22	0 0 151.3 21.8 16
23	1 0 13555.3 1588.5 16
24	2 0 165.3 21.7 16
25	3 0 13707.5 1831.6 16
26	4 0 89.0 12.4 16
27	5 0 141.3 16.4 16
28	6 0 11382.8 2062.2 16
29	7 0 170.5 19.3 16

Showing 9 to 29 of 507,452 entries

Figure 5. Visualization of GSE10072 CEL file

780	10.929071	10.022758	10.392993
5982	6.948867	6.815021	7.543393
3310	8.786485	8.098738	8.130723
7849	6.852304	6.894411	7.334652
2978	6.107564	6.170747	6.501881
7318	8.861774	9.404475	8.452138
7067	6.983388	7.112010	7.155872
11099	5.962803	6.282109	6.023964
6352	8.848925	8.733786	9.060864
1571	5.512906	5.725324	5.873940
2049	7.807251	7.742373	8.403955
2101	7.264559	7.334280	7.780656
1548	6.808536	7.048484	7.391876
100133684	9.825384	10.558581	8.326810
4323	8.399522	7.789321	8.517490
8717	8.154543	7.760409	7.751587
2342	7.949986	7.277906	8.318296
5337	5.627469	5.742076	5.787345
441263	8.888833	9.038015	9.336819

Showing 1 to 20 of 13,787 entries

Figure 6. Visualization of GSE10072 after data pre-processing

By comparing between Figure 2 and Figure 3, we can clearly differentiate the data arrangement whilst visualizing it. As in Figure 2, the dataset is arranged in 2 rows and the only way to differentiate the values are the spacing applied. There are also unwanted attributes such as patient biological information, dataset's information and so on. In Figure 3, the dataset is arranged in sequences and more rows are applied to differentiate between attributes. The right attributes are playing an important role in the algorithm because it can ease the running process of algorithm.

4. CONCLUSION

For cancer classification purpose, we presented the data pre-processing stage with the example of gene expression dataset. Gene expression dataset contains many attributes and much biological information about the samples. Hence, choosing the right attribute that could affect the algorithm is one of the important key steps which should not be ignored or neglected. Data pre-processing stages allow researchers to craft the dataset as intended. In this stage, the data will be accordingly clean and turned into the type of clean data that

is needed for the next machine learning process. For instance, we only focus on the selection of attributes that are weight-related.

Another direction of future research is to combine the cleaned data and enhanced the number of samples with a combination of other similar datasets. The number of samples in cleaned data after pre-processing stage is lower compared to the original dataset. Hence, it is possible to combine with other similar datasets to produce more samples in a dataset. It is a commonly seen scenario whereby there is a multitude of biological datasets which share the same features but were collected by different researchers under different experimental conditions. Though they may display different underlying distributions, they share highly relevant information. Each cleaned dataset contains limited samples, but high dimensions of gene expression value is insufficient to be considered as a good classifier. In such cases, transfer learning is one of the possible way to borrow more samples between datasets.

ACKNOWLEDGEMENTS

We would like to thank the Universiti Tun Hussein Onn Malaysia, Centre For Graduate Studies and Ministry of Higher Education Malaysia for supporting this research under the MYBRAIN15 and Fundamental Research Grant Scheme (Vot numbers: 1559). This paper was partly sponsored by the Centre for Graduate Studies UTHM.

REFERENCES

[1] Bair, E. Identification of significant features in DNA microarray data. *Wiley Interdisciplinary Reviews: Computational Statistics.* 2013;5(4):309-325.
[2] Kasim, S., Fudzee, M. F., Salamat, M. A., Ramli, A. A., Mahdin, H., & Abdullah, M. H. *An improved computational framework using one stage filtration by incorporating knowledge in gene expression clustering.* Proceedings of the International Conference on Artificial Intelligence and Robotics and the International Conference on Automation, Control and Robotics Engineering - ICAIR-CACRE 16. 2016.
[3] Sevugapandi, N. and Chandran, C. Classification Algorithm for Gene Expression Graph and Manhattan Distance. *Indonesian Journal of Electrical Engineering and Computer Science*. 2017;5(2):472.
[4] Liu W, Li C, Xu Y, Yang H, Yao Q, Han J et al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics.* 2013;29(17):2169-2177.
[5] Choon Sen, S., Kasim, S., Md Fudzee, M., Abdullah, R. and Atan, R. Random Walk From Different Perspective. *Acta Electronica Malaysia*. 2017;1(2):26-27.
[6] Seah C, Kasim S, Mohamad M. Specific Tuning Parameter for Directed Random Walk Algorithm Cancer Classification. *International Journal on Advanced Science, Engineering and Information Technology.* 2017;7(1):176.
[7] Revathy N, Amalraj D. Accurate Cancer Classification Using Expressions of Very Few Genes. *International Journal of Computer Applications*. 2011;14(4):19-22.
[8] Landi M, Dracheva T, Rotunno M, Figueroa J, Liu H, Dasgupta A et al. Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. *PLoS ONE*. 2008;3(2):e1651.
[9] Dai, Y., Guo, L., Li, M. and Chen, Y. Microarray Я US: a user-friendly graphical interface to Bioconductor tools that enables accurate microarray data analysis and expedites comprehensive functional analysis of microarray results. *BMC Research Note.* 2012;5(1):282.
[10] Odeh, A. Novel Genetic Algorithm for Early Prediction and Detection of Lung Cancer. *Journal of Cancer Treatment and Research*. 2017;5(2):15.
[11] Graudenzi, A. Pathway-based classification of breast cancer subtypes. *Frontiers in Bioscience*. 2017;22(10):1697-1712.
[12] Seah, C., Kasim, S., Fudzee, M., Law Tze Ping, J., Mohamad, M., Saedudin, R. and Ismail, M. An enhanced topologically significant directed random walk in cancer classification using gene expression datasets. *Saudi Journal of Biological Sciences.* 2017;24(8):1828-1841.
[13] Seah C, Kasim S, Fudzee M, Mohamad M. A Direct Proof of Significant Directed Random Walk. *IOP Conference Series: Materials Science and Engineering.* 2017;235:012004.
[14] Wu, J. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. *Biostatistics and Biometrics Open Access Journal.* 2017;1(2).
[15] Li, J., Meng, X., Wen, J. and Xu, Y. An Improved Method of SVM-BPSO Feature Selection Based on Cloud Model. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2014;12(5).
[16] Shi Jing, L., Fathiah Muzaffar Shah, F., Saberi Mohamad, M., Moorthy, K., Deris, S., Zakaria, Z. and Napis, S. A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data. *Current Proteomics*. 2015;12(1):14-27.
[17] Kim, Y. and Yoon, Y. A genetic filter for cancer classification on gene expression data. *Bio-Medical Materials and Engineering*. 2015;26(s1):S1993-S2002.