ORIGINAL RESEARCH

Submitted: 12.10.2022; Accepted: 8.11.2022; Published online: 25.12.2022

Gene Data Analysis for Disease Detection Using Data Mining Algorithms

Ramakrishnan Raman[1*]

[1]Symbiosis Institute of Business Management, Symbiosis International (Deemed University), Pune, Maharashtra, India
Email: director@sibmpune.edu.in*

Abstract

As a result of these promising results, researchers believe that gene expression tests are more important in creating more accurate and efficient diagnostic and classification tools for cancer. In the deoxyribonucleic acid (DNA) system, a gene is transcribed over ribonucleic acid (RNA) during transcription, which is a process known as gene expression (RNA). The study of gene expression data for cancer categorization has recently emerged as an active research subject. This research uses genetic algorithms (GA) to pick a subset of cancer microarray data that contains a meaningful set of genes. Then, standard classifiers like One-R, Bayesian Network, logistic regression, and Support Vector Machine (SVM) are developed based on these specific genes. Gene expression data sets are used to test the performance of these classifiers. According to the results of experiments, combining the confluence of GA and SVM is the most effective approach. In addition, the GA selection process is repeatable.

Keywords

Gene Data, Data Mining, RNA, DNA, Classification, Genetic Algorithm.

Imprint

Ramakrishnan Raman. Gene Data Analysis for Disease Detection Using Data Mining Algorithms. Cardiometry; Special issue No. 25; December 2022; p. 178-181; DOI: 10.18137/cardiometry.2022.25.178181; Available from: http://www.cardiometry.net/issues/no25-december-2022/gene-data-analysis

1. Introduction

Gene expression is the process of converting the deoxyribonucleic acid (DNA) sequence of a gene into the ribonucleic acid (RNA) sequence of the same gene. When a gene's expression level is high, it means that the RNA of that gene is generated in large quantities in a cell, and this is linked to the amount of protein produced by that gene. The advent of DNA microarray technology, which can simultaneously measure thousands of genome-wide expression values, has made it feasible to control gene sequencewithrecordcondition [1]. On this microarray, a glass slide is coated with a layer of single-stranded DNA molecules.

At least a few hundred thousand dots can be seen on an array, each representing a different gene. When it comes to using microarray technology, data analysis and management is becoming a serious challenge [2]. To compare the levels of messenger RNA (mRNA) in two samples, a microarray experiment is commonly utilised (e.g., treatment vs. control). Two separate fluorescent labels are used to label the RNA taken from the treatment and control cells. For example, the treatment cell RNA is labelled with red dye and the control cell RNA with green dye. Both extracts are applied on the microarray and allowed to dry [3].

The spots contain a corresponding sequence to the gene sequences in the extracts [4]. A laser is used to activate the array in order to determine the relative abundance of the hybridised RNA. It will be red if the treatment population's RNA is plentiful; it will be green if the control population's RNA is abundant. If both medication as well asregulatorpredicament evenly, the point would glow differently, however uncertaintywithbind, this would look black [5].

As a result, the relative increased expression of the proteins in the control and treatment populations may be calculated from the fluorescence intensities and colours for each site [6]. A general gene expression analysis by data mining is depicted in Figure 1.

2. Existing Work

The advent of DNA microarray technology, which can simultaneously measure thousands of genome-wide expression values, has made it feasible to visualize genetic factor countenance with record condition [7]. On this microarray, a glass slide is coated with a layer of single-stranded DNA molecules. At least a few hundred thousand dots can be seen on an array, each representing a different gene. When it comes to using microarray technology, data analysis and management is becoming a serious challenge. Biochemical conditions are taken into consideration when gene expression data is analysed [8]. A diagnosis

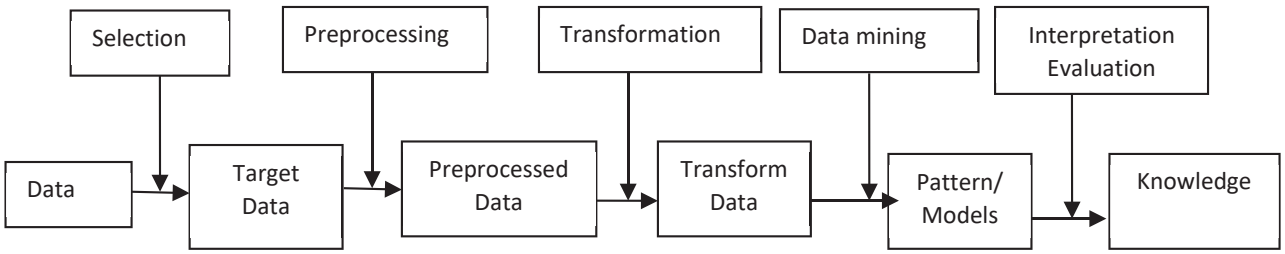

Figure 1. Data Mining Process for Gene Data Analysis

of cancer is a good illustration of this concept since the levels of expression of certain genes have been shown to change when cancer is present [9]. According to current research on tissue categorization at the molecular level, gene expression tests would play a vital role in the creation of efficient cancer diagnostic and classification platforms. The linear exclusionary study, proportional elective of data genetic factor, AI techniques, k-NN, clustering algorithms, and classifiers are only a few of the many suggested cancer classification methods [10].

The speed of calculation is also looked into, the accuracy of the categorization, and the capacity to uncover physiologically relevant gene information when evaluating the techniques multiclass classifiers for tissue categorization using gene expression [11]. When it came to multiclass gene expression data sets, several feature selection approaches and state-of-the art classification algorithms were tested. In comparison to the binary classification problem, the research shows this kind problem is substantially furtherhard to solve [12].

This included an examination of various important algorithms for multicategory classification, as well as gene selection approaches, numerous ensemble classification strategies, and two cross-validation designs on 11 datasets. Using gene expression data, for making accurate cancer diagnoses Multi-class SVMs are the greatest reliable ones was discovered [13]. Due to the large number of genes compared to training samples, this task is quite difficult. As a result, in order to cope up with colossal number of genes that aren't important this has to cope up with. In addition, the existence of distortions with dataset creates it is more difficult to accurately classify data when the sample size is insufficient [14].

3. Proposed System

Chromosome programming and fitness evaluation, selection, crossover, and mutation are major GA processes for gene selection using gene expression data. An initial vector representation is used to write the genome [15-16]. In order for a chromosome to survive and create kids, it must have the best suited chromosomes in the population [16]. Chromosome quality is improved, resulting in more accurate solutions to the optimization issue. A biased roulette wheel is used to choose parents from the population, and individually gene in the group comprises opening that is enlarged based on to its suitability. The proposed system architecture is shown in Figure 2.

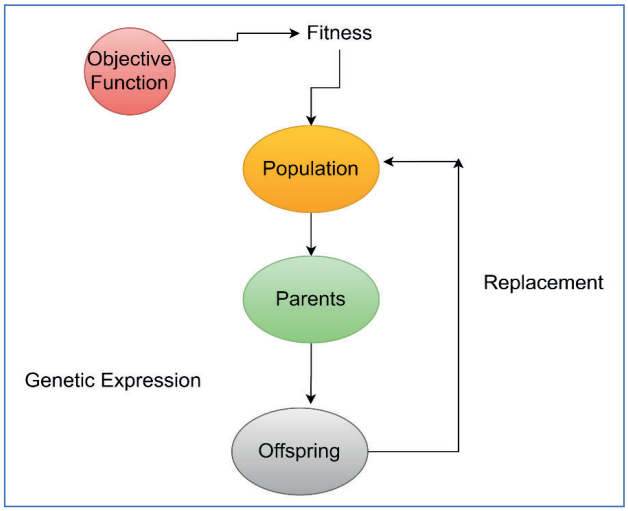

Figure 2. Proposed System Architecture

The crossover process is used to exchange information between two chromosomes that have been carefully selected. The two-point gene exchange is employed in this study. The points of intersection are determined at random. It is then completed by swapping genes with each other based on the crossover point in order to produce new children from those genes. Mutation happens following a crossover procedure. This is to avoid a local optimum of all solved issues being reached by all solutions in the population. Random genetic changes are made to new children in this study. Using a set of criteria, OneR constructs a single-level decision tree for testing a single gene.

A single gene is tested, and the results are compared to the rest of the population. The gene's value is represented as a branch, with each branch representing a distinct value. Using the training data, it is ideal to classify each branch according to the class that appears the most frequently. It's thus simple to establish how many instances of the rules are not containing-commongroup. A new class of guidelines is generated for each gene's cost, with one rule for each value. Select the gene whose rule set has the lowest mistake rate.

The root node of a decision tree is the highest one in the hierarchy. These tests are linked to a criterion for separating data sets into subgroups with improved class separability, which reduces the chance of misclassification. Heuristically pruning the tree to avoid overfitting, which might introduce performance of the classifier on the test data, is done after the tree has been formed from the training data.

4. Results and Discussion

While constructing the tree, this function is used to determine which gene should be divided, as well as which splitting criteria should be used for that gene. Using the purity entropy function, this is done by evaluating each unused gene at all of its possible split sites and picking the one that yields the best result. Six publicly accessible genomic data sets were taken from the literature to examine the effectiveness of the cascade of GA and various classifiers. Classifier based accuracy analysis is listed in Table 1.

Table 1

Classifier based Accuracy Analysis

Classifier	Accuracy
SVM	0.8457
GA	0.8540
Naïve Bayes	0.79568
One R	0.79654

All of them were normalised using the min–max method first. Then, a set of important genes was narrowed down using GA. Classifiers based on these genes were then built upon. Tenfold cross-validation involved hiding single fold of data while using samples from the other folds to create a decoder which might anticipate the hidden sample's classification. The accuracy and pace of change can be measured.

The entire study was conducted on a personal computer (PC) equipped with Microsoft Windows 8, an Intel Pentium 4 processor running at 2.0 GHz, and one gigabyte of random access memory (RAM). The J2SDK 1.6.0 development environment was utilised for the Java programming. The University of Waikato in New Zealand created the Week-3–4 data mining programme. SVM and DT were practically indistinguishable in terms of performance. In the case of a limited number of genes, DT performed better than SVM. In Figure 3, the bar chart for accuracy analysis has been depicted.

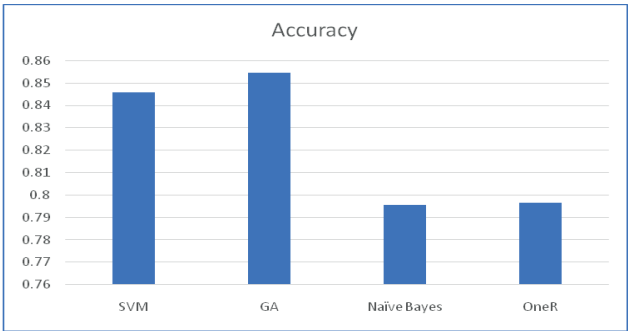

Figure. 3 Accuracy Analysis

In other words, SVM's performance was superior to that of DT's. GA and SVM have an accuracy of 88.84 percent (0.03), GA and Bayesian Network 72.11% (1.04), GA and OneR78.15% (1.02), and DT, GA have an accuracy of 88.80% (1.03). Regardless of the value of d, the patterns of gene selection are the same. Because of this, the GA selection pattern is stable. Only gene expression levels and Z-scores for the Colon information can be found here for the sake of brevity and page restriction. Number of genes-based accuracy analysis is shown in Figure 4.

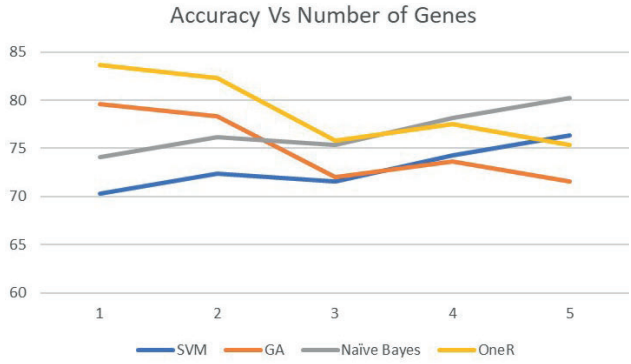

Figure 4. Number of Genes-based Accuracy Analysis

Whenever the number of genes picked was fewer than 20, SVM had a poor performance. In spite of SVM's inability to perform well with a small number of genes picked, the overall accuracy of SVM outperformed that of other methods. In Table 2, based on the number of genes, the accuracy values are listed.

Table 2

Accuracy based on Number of Genes

Number of Genes	SVM	GA	Naïve Bayes	One R
10	70.325	79.635	74.083125	83.625875
20	72.35	78.356	76.15875	82.3149
30	71.568	72	75.3572	75.8
40	74.258	73.65	78.11445	77.49125
50	76.35	71.53	80.25875	75.31825

5. Conclusion

One of the leading causes of mortality has been cancer for a long time now.There is presently no reliable preventative medicine available. The authors of this research use GA to narrow down a large set of cancer microarray data to a manageable set of important genes. Those genes are then used to build popular classifiers. Gene expression data sets are used to test the performance of these classifiers. Experiments show that the combination of GA with SVM is the most effective strategy for analysing data. When there are more than 30 genes involved, the accuracy of this method remains consistent. GA's gene selection process is repeatable, since few genes remainreliably preferred withevery dataset with different values. In the realm of clinical science, the gene expression data set of lung malignant development is particularly relevant. Characterization and determination procedures are critical in accurately identifying a condition, making the analytical process and right diagnosis easier. As the number of highlights or features was reduced, the characterisation approaches showed modest improvement.

References

1. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., et al. 2000. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature, 403(6769):503–511.
2. Bolshakova, N. and Azuaje, F. 2003. Cluster validation techniques for genome expression data. Signal Processing, 83:825–833.
3. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., and Haussler, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences USA (PNAS), 97(12):262–267
4. Chu, F. and Wang, L. 2005. Applications of support vector machines to cancer classification with microarray data. International Journal of Neural Systems, 15(6):475–484.
5. Dudoit, S., Fridlyand, J., and Speed, T. P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 97(457):77–87.
6. Golub, T. R., Slonim, D. K., Tamayo, P., Gaasenbeek, M., Huard, C., Mesirov, J. P., Coller, H., et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531–537.
7. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. Machine Learning, 46: 389–422.
8. Amalarethinam, DI George, and N. Aswin Vignesh. "Prediction of diabetes mellitus using data mining techniques: a survey." International Journal of Applied Engineering Research 10, no. 82 (2015): 2015.
9. Liu, J., Iba, H., and Ishizuka, M. 2001. Selecting informative genes with parallel genetic algorithms in tissue classification. Genome Informatics, 12:14–23.
10. Jaiganesh, V., M. Tech, S. Murugan, and Hardware Design Engineer. "PC Based Heart Rate Monitor Implemented In Xilinx Fpga And Analysing The Heart Rate." In Circuits, Signals, and Systems (pp. 319-323).
11. Statnikov, A., Aliferis, C. F., Tsamardinos, L., Hardin, D., and Levy, S. 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21(5):631–643.
12. Amalarathinam, DI George, and N. Aswin Vignesh. "A New Monotony Advanced Decision Tree Using Graft Algorithm to Predict the Diagnosis of Diabetes Mellitus." International Journal of Pure and Applied Mathematics 118, no. 6 (2018): 19-28.
13. Yip, K. Y., Cheung, D. W., and Ng, M. K. 2004. HARP: A practical projected clustering algorithm. IEEE Transaction on Knowledge and Data Engineering, 16(11):1387–1397.
14. Zhang, H., Yu, C. Y., and Singer, B. 2003. Cell and tumour classification using gene expression data: Construction of forests. Proceedings of the NationalAcademy of Sciences USA (PNAS), 100(7):4168–4172.
15. G. Prakash. (2019). Deduplication with attribute based encryption in E-health care systems. International Journal of MC Square Scientific Research. 11(4). pp: 16-2483.
16. Balamurugan E, Akpajaro J. Genetic Algorithm With Bagging For DNA Classification. International Journal of Advances in Signal and Image Sciences. 2021;7(2):31-39.