Classification of Cancer Gene Subtypes from Clustering of Gene Expression Data

Logenthiran Machap, Afnizanfaizal Abdullah, Zuraini Ali Shah

Abstract: ***Typically, microarray gene expression data obscure imperative information which is necessary for the understanding of molecular biology processes that occurs in a specific organism with respect to its environment. Uncovering gene expression data's invisible patterns will lead to a remarkable desire to enhance the interpretation of functional genomics. Biological networks intricacy and the presence of huge amount of genes raise the difficulties of understanding of the high dimension data, which resides lots of measurements. Thus, clustering techniques which are crucial in the data mining process are used as the first step to address this challenge to discover logical structures and predict significant patterns in the hidden data. These patterns may offer shreds of evidence about the biological process related to different physiological conditions. On deep, this paper focuses on the co-clustering algorithm to cluster genes and conditions simultaneously to obtain co-clusters further utilised for classification. The method called an improved network assisted co-clustering for the identification of cancer subtypes (iNCIS). Fundamentally, it integrates gene network information with gene expression to achieve biologically significant clusters. The classes obtained from clusters were used in the classification of genes to improve accuracy. This method applied to breast cancer and glioblastoma multiforme datasets. The discovered structures disclosed strong biological significance associations between functional annotations of genes with related conditions.***

Index Terms: ***classification, clustering, gene expression, microarray***

I. INTRODUCTION

For decades, microarray technology has arisen as an effective technique to measure the expression levels of thousands of genes with various conditions. Molecular biologists have a key interest to detect genes expression levels which change in different experimental conditions where it is significant to understand gene regulation, gene function, cellular processes, gene subtypes and many more. However, for most of the available microarray datasets, the number of samples is small compared to thousands of genes [1]. It is essential to reduce the data for gene expression analysis which is necessary for advanced analysis such as clustering and classification [2, 3].

Scientists have been applied numerous unsupervised learning of clustering analysis techniques in gene expression matrix and these techniques successfully identified biologically significant clusters of genes and samples [4]. Clustering is a part of data mining from the machine learning area [5, 6]. Some examples of traditional clustering algorithms such as hierarchical clustering [7], k-means [8], and self-organizing maps (SOM) [9] were implemented in gene expression data although it was developed for non-biological research. Conversely, most of these techniques have failed to identify co-regulated and co-expressed gene subsets under different experimental conditions.

Therefore, co-clustering was replaced with traditional clustering to uncover local coherent patterns which reveal better biological certainty by simultaneously clustering the genes and samples. Hartigan [10] is the first who implemented the co-clustering algorithm known as direct clustering for data matrix. Researchers did an enormous extent of work on co-clustering algorithms in numerous perspectives. Co-clustering was applied different domain, such text mining [11], search in databases [12], and target marketing [13].

While Cheng and Church [14] are applied co-clustering also called ad bi-clustering techniques for gene expression data specifically. Greedy search heuristic to obtain co-clusters using mean squared residue value. On top of that, Kluger, Basri [15] adopt spectral bi-clustering algorithm from Dhillon [16] on gene expression data to generate structure like 'checkerboard'. Beside this, an information-theoretic co-clustering algorithm was proposed by Dhillon, Mallela [17] to utilize non-negative matrices.

Recently, Fuzzy Co-clustering algorithm based on information bottleneck similarity known as ibFCC proposed by Liu, Wu [3] using similarity measure. Essentially, this method assigns membership functions of genes and samples; hence it was applied to biomedical data. There are some review papers also available for co-clustering algorithms which can be referred for more understanding of current trends in these algorithms [18-20]. Figure 1 shows the general taxonomy division of co-clustering techniques for gene expression analysis. Thus, in this paper, it has been proposed for classification of gene subtypes generated from unsupervised learning of an improved co-clustering algorithm.

Revised Manuscript Received on May 05, 2019.

Logenthiran Machap, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia , Johor, Malaysia.
Afnizanfaizal Abdullah, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia , Johor, Malaysia.
Zuraini Ali Shah, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia , Johor, Malaysia.

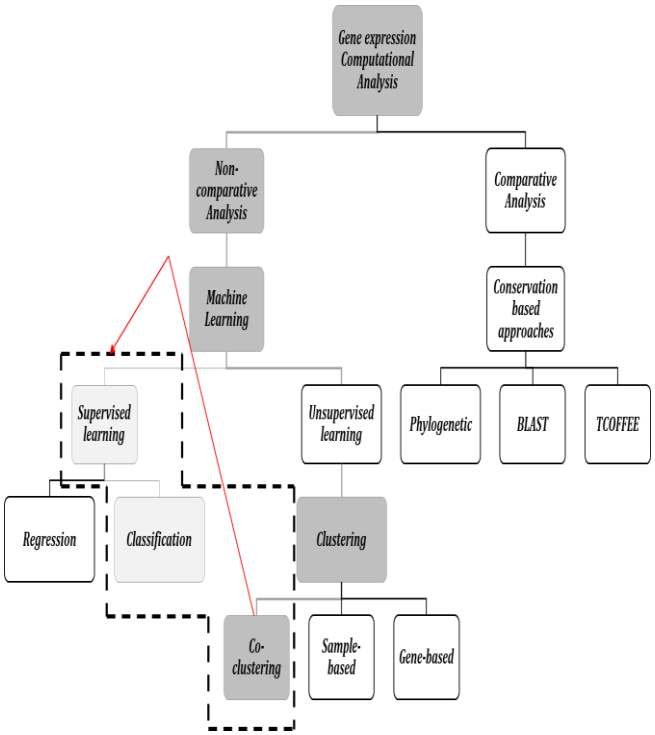

Fig. 1 Taxonomy overview of gene expression analysis

Therefore, unsupervised learning by clustering possible to use as a pre-processing step for supervised learning known as classification. Categorizing the genes into their respective class as normal or a tumour accurately will be done using the classifier. Classifier is known as an artificial intelligence scheme to make the potential prediction. Generally, in cancer identification layout, most of the existing algorithms are focusing on obtaining higher prediction accuracy even though the class size is different and uneven. Examples of classifiers are support vector machine (SVM), neural network (NN), k-nearest neighbour (kNN) and classification tree [21]. Figure 1 dashed line shows the general view on classification from clustering.

II. MATERIALS AND METHODS

Primarily, cancer gene expression dataset was downloaded from publicly available the cancer genome atlas (TCGA) project. The method applied in this paper is about the classification of gene subtypes after clustering the data. And the algorithm was implemented in MATLAB. The experimental setup for this research involves MATLAB programming language and the development platform is Windows environment. The MATLAB script is implemented with NCIS and SVM for conducting this experiment. This experiment is conducted using a computer with 16GB of RAM, Intel core i7-4790 CPU with 3.60GHz, and 1TB storage size.

A. Gene expression data

In general, two cancer microarray dataset with ***n*** genes in ***m*** samples were used in this research. They are Breast Cancer (BRCA) and Glioblastoma Multiforme (GBM) datasets. In these both datasets, the number of genes is higher than the number of samples. Essentially, all publicly available gene expression matrices have been pre-processed in several ways, such as image analysis, expression quantization, normalization, and screening out. Table I shows the original datasets were downloaded from the online publicly available database.

Table. 1 Datasets

Dataset	# Genes	# Samples	Reference
Breast cancer	17814	547	[22]
Glioblastoma Multiforme	11861	202	[23]

B. Clustering and Classification

Author Liu, Gu [24] first proposed the gene subtypes co-clustering algorithm called Network assisted Co-clustering for the Identification of cancer Subtypes (NCIS). Researcher utilises Semi-Nonnegative Matrix Tri-Factorization (SNMTF) method from matrix factorization-based clustering family in this algorithm. Their main idea is to incorporate biological information such as gene network information with gene expression data in the clustering process. Undeniably in cancer, network information is significant to comprehend the molecular complexity [25, 26]. Most existing methods did not integrate molecular interaction information. Incorporation of network information will lead to produce more knowledge of interaction with system level besides improving the capability to identify cancer subtypes.

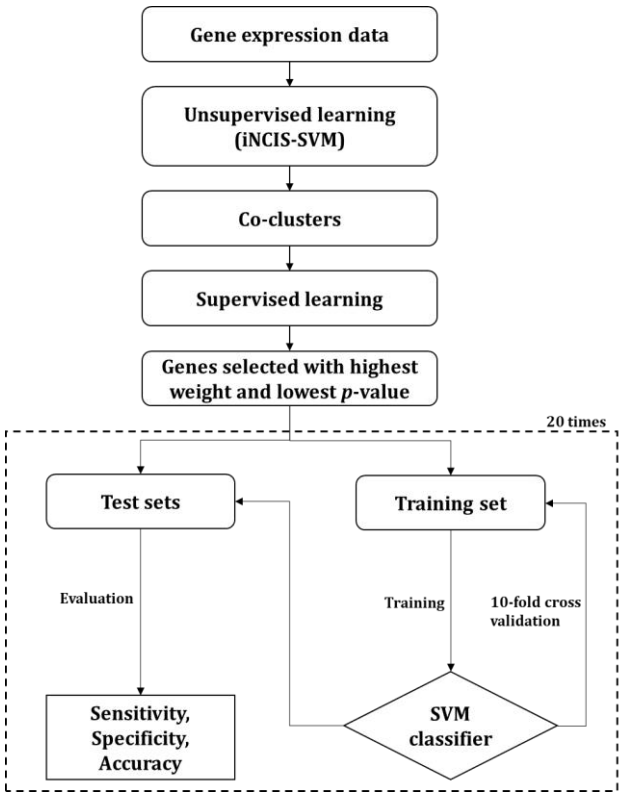

Fig. 2 The process of proposed classification after clustering

As our first objective, is to propose an improvement in NCIS (iNCIS) algorithm in the objective function section. The first step in this algorithm is assigning weights for each

gene from the gene expression matrix. Beside this, gene network information was integrated with gene weights. Finally, the significant gene will receive higher weights and will be measured further in the weighted co-clustering algorithm to obtain co-clusters of subtypes. This paper, particularly presents our second objective which is the classification of cancer subtypes identified from the first objective where iNCIS adopted with support vector machine (SVM). Principally, clustering is unsupervised learning which is divided into two: soft and hard clustering. In hard clustering, variable from dataset fits precisely one cluster while in a soft clustering variable assigned to all the clusters with different probabilities or weights [27]. Through the improvement in the co-clustering technique provides the flexibility and options to generate different gene subtypes with better accuracy beside adapt to the classifier to obtain optimal classification accuracy of cancer gene identification. Figure 2 shows the general process of the proposed classification objective.

III. RESULTS AND DISCUSSION

Breast cancer and glioblastoma datasets are used in this research. From the raw datasets, modified PageRank algorithm was applied to assign weights to genes in both datasets. Table II shows the number of genes selected from the PageRank algorithm. Weight is assigned to each gene in the dataset by integrating with network information which was collected from Reactome, the NCI-curated PID and KEGG. That network information is gene co-expression, protein interaction, and protein domain interaction.

Table. 2 Selected genes after assigns weight by PageRank algorithm

Dataset	# Genes	# Samples	Reference
Breast cancer	8726	547	[22]
Glioblastoma Multiforme	7183	202	[23]

These selected genes, then directly applied in the co-clustering algorithm to obtain cancer subtype with class. Through co-clustering, we able to obtain five subtypes of breast cancer and four subtypes for glioblastoma multiforme datasets. Table III shows the results from co-clustering. From the clustering process, we able to attain the classes of samples to be further used for the classification process. Since classification is a supervised learning method, the classes obtained from unsupervised learning of clustering in the first objective are continuously utilised in the classifier.

Table. 3 Class of datasets from clustering process

Dataset	# Genes	# Samples	Class
Breast cancer	8726	547	5
Glioblastoma Multiforme	7183	202	4

The complete analysis for the selected cancer datasets had been tabulated according to a selected range of selection both the number of genes in subtypes and accuracy rate. The results obtained are an averaged from multiple runs with a total iteration of 20 runs per selected gene range for each dataset. The selected gene range had been set to 35, 50, 100, and 200. This selection was done by performing ANOVA test for each gene expression level across the five subtypes. From the test, we select the genes with the largest weights and smallest p-value. We performed 10-fold cross validation.

In order to verify the effectiveness of the proposed method, we compared it with the original NCIS and NetBC methods of default parameters in MATLAB. Beside, SVM used in this experiment is derived from the LIBSVM package developed by Chang and Lin [28]. On top of that, since the numbers of the selected genes are small, 10-fold cross validation was applied. In addition, to avoid unstable operation results, each experiment was run 10 times and then averaged classification accuracy was calculated for comparison. The final classification results are shown in table IV.

Table. 4 Prediction accuracy for various gene selections range and accuracy comparison

Method / Genes	iNCIS+SVM		NCIS+SVM	
	BRCA	GBM	BRCA	GBM
35	**0.936**	**0.905**	0.901	0.902
50	**0.960**	**0.931**	0.932	0.903
100	**0.945**	**0.926**	0.902	0.890
200	**0.946**	**0.916**	0.915	0.900

For our analysis, we can conclude that the suitable range or amount of significant genes was 50, where both datasets have shown better or higher accuracy in this selected number of genes. Although the accuracy difference was not irregular, the selected numbers of gene were either too less or too many compared to other selections. However, other selection variance of genes can be used by other researchers for consequent analysis besides filtering of genes for huge datasets. In between, the selection of genes can be varied according to requirements for example, gene network analysis, gene ontology for gene functional annotation many more.

The justification for the improvement achieved from this classification subsequent from clustering. A comparison with previous work was done according to the accuracy is shown in figure 3. From the comparison, can conclude that overall our method has raised the classification accuracy for both datasets used. This improvement was achieved because the improvement made in the clustering process help us to select more significant genes compared to previous research.

From Table IV, it has been concluded that the classification accuracy is achieved more than 93% for BRCA data set and more than 90% for GBM data set. The comparison was made for the same number of selected genes from both data sets. The range of accuracy and error rate falls between 0 and 1. If the accuracy is close or equal to 1 means it is a better classification.

The number of genes selected for both data sets is 50, which shows the highest accuracy compared to the other selected number of genes. BRCA shows 3% of accuracy increment compared to the previous method while GBM dataset show 2.99% accuracy increment for 50 genes.

In addition, the proposed method shows improvement on all the selected data sets as shown in Table IV. Somehow, the accuracies shown were not very stable for the genes number 100, and 200. Hence, 50 genes are considered as the best number of genes for this proposed method.

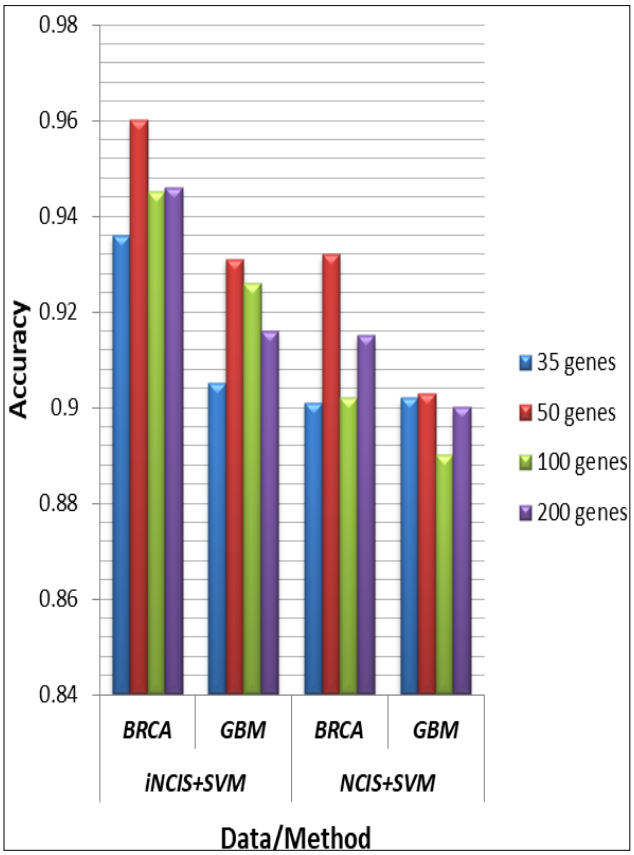

Fig. 3 Prediction accuracy comparison

Figure 3 is the bar graph for both data sets according to the assorted number of genes as in Table IV. For the 35 genes, BRCA shows 3.88% improvement whereas GBM shows a very small fraction of improvement which is 0.33%. Besides this for 100 genes, BRCA produced 4.78% enhancement while GBM generated 4.04% increment of accuracy. On top of that, for the 200 genes from BRCA produced 3.39% increment. However, for GBM, the proposed method show rise on accuracy about 1.78% on 200 selected genes. Essentially, different number of genes selected to train classifier and obtained the accuracy approximations based on the calculation of average accuracy of 10-fold cross validation for 20 times.

IV. CONCLUSION

Gene expression classification of patient samples has been focused on cancer diagnosis and prognosis. To deal with the high dimensionality of data and low classification accuracy our method outperforms the previous research methods of identifying biologically significant genes. These genes justification analysis carried out further in terms of the functions of genes and pathway analysis. From the research, it has been noticed that the default parameter setting was adopted for both of the datasets. Using a different parameter setting also, the experiment was conducted, but the results weren't significant. The gene list was selected according to the weight that was assigned to each gene. If possible, additional gene selection technique could be applied in order to obtain optimal gene sets for advance level experimental analysis. Additional analysis needs to be carried out on selected gene functions and pathway prediction from Gencards and DAVID. Future work can be done by applying the algorithm to single nucleotide polymorphism (SNP) datasets will be essential in cancer gene identification.

ACKNOWLEDGMENT

The author is grateful for the guidance of the supervisor and co-supervisor. Besides this, would like to thank the members of the Synthetic Biology Research Group and Artificial Intelligence and Bioinformatics Group from Universiti Teknologi Malaysia for useful discussions that make this paper successfully completed.

REFERENCES

1. Yang, S. and D.Q. Naiman, *Multiclass cancer classification based on gene expression comparison.* Statistical applications in genetics and molecular biology, 2014. **13**(4): p. 477-496.
2. Liu, W., K. Yuan, and D. Ye, *Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis.* Journal of Biomedical Informatics, 2008. **41**(4): p. 602-606.
3. Liu, Y., et al., *A fuzzy co-clustering algorithm for biomedical data.* PLOS ONE, 2017. **12**(4): p. e0176536.
4. Cho, H. and I.S. Dhillon, *Coclustering of Human Cancer Microarrays Using Minimum Sum-Squared Residue Coclustering.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2008. **5**(3): p. 385-400.
5. Abdullah, A., et al. *An improved local best searching in Particle Swarm Optimization using Differential Evolution*. in *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*. 2011.
6. Abdullah, A., et al., *An improved swarm optimization for parameter estimation and biological model selection.* PLoS One, 2013. **8**(4): p. e61258.
7. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences, 1998. **95**(25): p. 14863-14868.
8. Ghosh, D. and A.M. Chinnaiyan, *Mixture modelling of gene expression data from microarray experiments.* Bioinformatics, 2002. **18**(2): p. 275-286.
9. Map, S.-O. and T. Kohonen, *Self-organizing map.* Proceedings of the IEEE, 1990. **78**: p. 1464-1480.
10. Hartigan, J.A., *Direct Clustering of a Data Matrix.* Journal of the American Statistical Association, 1972. **67**(337): p. 123-129.
11. Lewis, D.D., et al., *Rcv1: A new benchmark collection for text categorization research.* Journal of machine learning research, 2004. **5**(Apr): p. 361-397.
12. Agrawal, R., et al., *Automatic subspace clustering of high dimensional data for data mining applications*. 1999, Google Patents.
13. Wang, H., et al. *Clustering by pattern similarity in large data sets*. in *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. 2002. ACM.
14. Cheng, Y. and G.M. Church, *Biclustering of Expression Data*, in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*. 2000, AAAI Press. p. 93-103.
15. Kluger, Y., et al., *Spectral biclustering of microarray data: coclustering genes and conditions.* Genome research, 2003. **13**(4): p. 703-716.
16. Dhillon, I.S. *Co-clustering documents and words using bipartite spectral graph partitioning*. in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 2001. ACM.

17. Dhillon, I.S., S. Mallela, and D.S. Modha, *Information-theoretic co-clustering*, in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, ACM: Washington, D.C. p. 89-98.
18. Padilha, V.A. and R.J.G.B. Campello, *A systematic comparative evaluation of biclustering techniques.* BMC Bioinformatics, 2017. **18**(1): p. 55.
19. Eren, K., et al., *A comparative analysis of biclustering algorithms for gene expression data.* Briefings in bioinformatics, 2012. **14**(3): p. 279-292.
20. Freitas, A., et al., *Survey on biclustering of gene expression data.* Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, 2013: p. 591-608.
21. Lin, W.-J. and J.J. Chen, *Class-imbalanced classifiers for high-dimensional data.* Briefings in bioinformatics, 2012. **14**(1): p. 13-26.
22. Network, C.G.A., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61.
23. Verhaak, R.G.W., et al., *Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
24. Liu, Y., et al., *A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression.* BMC Bioinformatics, 2014. **15**(1): p. 37.
25. Chuang, H.-Y., et al., *Network-based classification of breast cancer metastasis.* Molecular Systems Biology, 2007. **3**: p. 140.
26. Barabasi, A.-L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nat Rev Genet, 2011. **12**(1): p. 56-68.
27. Panda, B., S. Sahoo, and S.K. Patnaik, *A comparative study of hard and soft clustering using swarm optimization.* International Journal of Scientific & Engineering Research, 2013. **4**(10): p. 785-790.
28. Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM transactions on intelligent systems and technology (TIST), 2011. **2**(3): p. 27.

AUTHORS PROFILE

Logenthiran Machap received B.Sc. Bioinformatics from National University of Malaysia (UKM), Master of Computer Science from University of Technical Malaysia Melaka (UTeM). He is a PhD candidate of computer science from University of Technology Malaysia (UTM). He is currently working on co-clustering algorithm and its application on cancer microarray gene expression data. His research interests include data mining, machine learning, artificial intelligence and bioinformatics.

Afnizanfaizal Abdullah holds B.Sc., M.Sc., and PhD in computer science from University of Technology Malaysia (UTM). He's a Senior Lecturer at School of Computing in Faculty of Engineering at UTM. Besides this, he is a Deputy Director of UTM Centre for Student Innovation & Technological Entrepreneurship and served as Chair for IEEE Young Professionals (Malaysia Section). With background of computing and computer science, his proficiency involved around Data Science comprise Machine learning methods for huge data analysis; Computational Systems and Synthetic Biology in designing and improving biological data models.

Zuraini Ali Shah Zuraini Ali Shah is Lecturer of the Department of Software Engineering, School of Computing, Faculty of Engineering at UTM, where she has been since 2000. She received a B.Sc and M.Sc from UTM. She received her Ph.D. in Computer Science in 2012. Her research interests include Computational Intelligence in Pattern recognition, Machine learning in Bioinformatics and others