UNIVERSITI PUTRA MALAYSIA

INTEGRATION OF RANK-PARTITION AND SEQUENTIAL WRAPPER TECHNIQUES FOR FEATURE SELECTION OF BREAST CANCER MICROARRAY DATA

AHMED ABBAS ABDULWAHHAB

INTEGRATION OF RANK-PARTITION AND SEQUENTIAL WRAPPER TECHNIQUES FOR FEATURE SELECTION OF BREAST CANCER MICROARRAY DATA

By

AHMED ABBAS ABDULWAHHAB

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Master of Science

November 2015

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Master of Science

INTEGRATION OF RANK-PARTITION AND SEQUENTIAL WRAPPER TECHNIQUES FOR FEATURE SELECTION OF BREAST CANCER MICROARRAY DATA

By

AHMED ABBAS ABDULWAHHAB

November 2015

Chairman : Makhfudzah Binti Mokhtar, PhD
Faculty : Engineering

A deoxyribonucleic acid (DNA) microarray has the ability to record huge amount of genetic information simultaneously. Previous researches have shown that this technology can be helpful in the classification of cancers and their treatments outcomes. This has encouraged information technology engineers to cooperate in microarray data analysis for enhancing medicine and biology technologies .

Typically, cancer-related microarray data are consisted of high dimensional gene expression levels (as features) for a limited number of samples. This characteristic in the structure of microarray data causes the phenomenon known as the curse of dimensionality, which is a particularly problem for standard classification models. It contradicts to the required ratio of samples to genes which should be much greater than 1 and it makes the direct application of machine learning techniques inefficient. Consequently, gene selection techniques have become a crucial element in the classification of microarray data .

Based on previous researches in the context of microarray data classification, the results obtained from the classification of breast cancer data have the lowest accuracy among them. Therefore, this study was aimed in improving the classification accuracy of clinical outcomes for breast cancer by gene expression profiling .

Filter and wrapper for gene selection are the main techniques in many existing microarray data analysis. Promising results obtained from filter-wrapper techniques have led to the design of a proposed model for this study. A gene selection model that integrates rank-partition and sequential wrapper was designed to find optimal subset of the most informative genes that enhances the predictive power of gene expression profiling .

Evaluation of the obtained results for breast cancer data set demonstrates that the proposed integrated model achieved the objective in finding the optimal subset of the most informative genes that has the predictive power of 87% accuracy compared to 83% of the original study and 77% of the shrunken centroid method.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PENGINTEGRASIAN TEKNIK PENEMPATAN-PENYEKATAN DAN PEMBUNGKUSAN BERTURUTAN UNTUKPEMILIHAN SIFAT DALAM PENGKELASAN DATA SUSUNAN MIKRO

Oleh

AHMED ABBAS ABDULWAHHAB

November 2015

Pengerusi : Makhfudzah Binti Mokhtar, PhD
Fakulti : Kejuruteraan

Susunan mikro asid deoksiribonukleik (DNA) mempunyai keupayaan untuk merekodkan sejumlah besar maklumat genetik pada masa yang sama.Kajian terdahulu telah menunjukkan bahawa teknologi ini mampu membantu dalam klasifikasi kanser dan hasil rawatan.Ini telah menggalakkan jurutera teknologi maklumat untuk bekerjasama dalam penganalisaan data susunan mikro untuk meningkatkan teknologi perubatan dan biologi .

Biasanya, data kanser yang berkaitan dengan susunan mikro terdiri daripada tahap ekspresi gen (sebagai sifat) yang berdimensi tinggi bagi bilangan sampel yang terhad.Ciri-ciri ini menyebabkan fenomena yang dikenali sebagai 'laknat kematraan' dalam struktur data susunan mikro yang menjadi masalah terutamanya untuk model klasifikasi standard.Ia bercanggah dengan nisbah sampel kepada gen yang diperlukan yang sepatutnya jauh lebih besar daripada 1, dan ini menyebabkan penggunaan teknik pembelajaran mesin secara langsung menjadi tidak cekap.Oleh itu, teknik pemilihan gen menjadi elemen penting dalam pengelasan data susunan mikro .

Berdasarkan kajian dalam konteks klasifikasi data susunan mikro sebelum ini, keputusan yang diperolehi daripada klasifikasi data kanser payudara mempunyai ketepatan yang paling rendah di kalangan mereka.Oleh itu, kajian ini bertujuan untuk meningkatkan ketepatan klasifikasi hasil klinikal kanser payudara melalui profil ekspresi gen.Penapis dan pembungkus untuk pemilihan gen adalah teknik utama dalam banyak analisis data susunan mikro sedia ada.Kejayaan awal yang diperolehi daripada teknik penapis-pembungkus telah membawa kepada model reka bentuk yang dicadangkan untuk kajian ini.Model pemilihan gen yang mengintegrasi kaedah penapis(penempatan sifat), penyekatan sifat dan pembungkusberturutan telah direka untuk mencari subset gen optimum yang paling bermaklumat yang meningkatkan kuasa ramalan profil ekspresi gen .

Penilaian keputusan yang diperolehi untuk set data kanser payudara menunjukkan bahawa model integrasi yang dicadangkan mencapai objektif dalam mencari subset gen optimum yang paling bermaklumat yangmempunyai kuasa ramalan ketepatan 87% berbanding 83% dalamkajianasal dan 77% dalam kajian model pemusatan kecut .

ACKNOWLEDGEMENTS

Praise to **Allah the Almighty**, and peace be upon our **Prophet Mohammed**. At the beginning , I must thank **ALLAH SWT** for numerous blessings among which is the completion of this thesis.

I would like to take this opportunity to express my gratitude and appreciation towards my supervisor, **Dr. Makhfudzah Binti Mokhtar** , for the endless supervisory effort and time spent with me. Without her advice and guidance from time to time, this project could not have made progress and this thesis would never have come into existence. Also, I would like to express my grateful to my supervisory committee members **Professor Dr. M. Iqbal Bin Saripan** and **Dr. Muhammad Hafiz Bin Abu Bakar**. Thank you all.

Special thanks must go to my wife, **Muna** , who has always been there, providing unconditional love, support and understanding throughout my study.

I would like to express my grateful to my family all my friends in Iraq and Malaysia who have never stopped supporting me with every encouraging words and feelings.

I cannot forget to thank the Nederland Institute of Cancer NKI, who provided me with breast cancer dataset to pursue a M.Sc. project.

At last, but not the least, I ask **ALLAH SWT** to have mercy upon **Martyrs of Iraq** and their families, I want to express my love and wishes of peace and prosperity for my country (**Iraq**). And I would like to express my love and grateful to the beautiful country, **Malaysia**, for its fantastic hospitality.

I certify that a Thesis Examination Committee has met on 27th November 2015 to conduct the final examination of Ahmed Abbas Abdulwahhab on his Master of Science thesis entitled " Integration of Rank-Partition and Sequential Wrapper Techniques for Feature Selection of Breast Cancer Microarray Data " in accordance with Universiti Pertanian Malaysia (Higher Degree) act 1980 and Universiti Pertanian Malaysia (Higher Degree) regulations 1981. The Committee recommends that the candidate be awarded the Master of Science. Members of the Examination Committee are as follows:

This thesis was submitted to Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows :

Makhfudzah Binti Mokhtar, PhD
Senior Lecturer
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

M. Iqbal Bin Saripan, PhD
Professor
Faculty of Engineering
Universiti Putra Malaysia
(Member)

Muhammad Hafiz Bin Abu Bakar, PhD
Senior Lecturer
Faculty of Engineering
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institution;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to Universiti Putra Malaysia (Research) rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature:_____ Date: 23/3/2016

Name and Matric No: Ahmed Abbas Abdulwahhab GS35463

Declaration by Members of Supervisory Committee

This is to confirm that:

- this research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: _____

Name of

Chairman of

Supervisory

Committee: Dr. Makhfudzah Binti Mokhtar

Signature: _____

Name of

Chairman of

Supervisory

Committee: Professor Dr. M. Iqbal Bin Saripan

Signature: _____

Name of

Chairman of

Supervisory

Committee: Dr. Muhammad Hafiz Bin Abu Bakar

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic Acid
IG	Information Gain
SNR	Signal to Noise Ratio
FCCEGS	Fuzzy C-mean Clustering-based Enhanced Gene Selection
MGS_SOM	Microarray Gene Selection by using Self-Organizing Maps
LS Bound-SFS	Least Squares Bound combined with Sequential Forward Selection
SFGS	Sequential Forward Gene Selection
MIFS	Mutual Information Feature Selection
RFE	Recursive Feature Elimination
SC.s	Shrunken Centroids
PCA	Principle Component Analysis
NFE	Neuro-Fuzzy Ensemble
SVM	Support Vector Machine
KNN	K- Nearest Neighbor
BSVM	Biased Support Vector Machine
LS-SVM	Least Squares Support Vector Machine
GP	Genetic Programming
GATree	Genetically Evolved Decision Tree
RF	Random Forests
ANN	Artificial Neural Network
LOOCV	Leave-One-Out Cross-Validation
K-fold CV	K-fold Cross-Validation
AUC	Area Under Receiver-Operating Characteristic Curve
NCI	National Cancer Institute

CHAPTER ONE

INTRODUCTION

1.1 Introduction

This thesis aims at reporting the work that has been done in the research study titled "Integration of Rank-Partition and Sequential Wrapper Technique for Feature Selection of Breast Cancer Microarray Data " for the degree of Master of Science. In addition to the literature review part which reconsiders relevant works previously carried out by other studies, the description of employed methodology and discussion on the analyzed results for the outcomes of this study, this thesis describes the proposed integrated model for achieving the objective of this study.

1.2 Background

The biological functions of life are performed through orchestrated manner interactions among huge number of genes and their corresponding proteins. As a result of technological advances, now there is another method for looking into disease itself in addition to conducting tedious molecular laboratory experiments [2]. In conventional molecular methods, one gene in an experiment is focused [18]. Therefore, they are not useful to construct the full image of the nature of interactions between gene-gene or gene-protein. In the last decades, the emergence of DNA microarray technology enhances the situation by providing the ability for monitoring the expression levels for tremendous amount of genes simultaneously. The technology of microarray has been thought to be able to offer a method of understanding the disease complexity in a molecular level, improving the accuracy of diagnostic of disease and specifying potential aims for prediction and therapies [2] .

1.3 Problem Statement and Motivation

During the recent decades, an enormous throughput of biological data extracted by microarray technology have been deposited in gene banks and the Internet for information retrieval and further research studies [1],[2]. Microarray experiments and their data analysis may assist in more full comprehension of the molecular variations among tumors and thus to more accurate and reliable classification through the monitoring of expression levels for huge amount of genes in cells simultaneously [18]. These genetic issues seem can be overcome more easily by using the techniques of machine learning. Microarray technology along with classification techniques has successfully guided the decisions of clinical management for individual patients, such as oncology [2]. Profiles of gene expression, which obtained from experiments of microarray, represent a snapshot of expression levels of up to tens of thousands of genes for limited number of samples. Analysis of such data , large gene-to-sample ratio , can be impractical in addition to this data may be occasionally noisy (i.e. most of these genes are not contributing in distinction between the classes of data). Increasing the size of samples is difficult to be achieved due to the following reasons; First, producing high-throughput gene expression data is extremely expensive. Second,

difficulties of persuading patients to join in the research studies. Third, some diseases are difficult to be morphologically distinct [2],[18] . This characteristic in the structure of microarray data causes the phenomenon known as the *curse of dimensionality* , where contradicts to the ratio of samples to genes which should be much more than 1. This phenomenon of high-dimensional genes is a particularly problem for standard classification models. Often a deterioration in performance is observed when classifying cancer- related microarray data. In order to overcome the problem of the *curse of dimensionality*, a subset of genes should be extracted from the entire set of genes in a process called gene selection. These genes are truly relevant to the status of disease and they known as the most informative genes [1],[2].

Based on previous research studies in the context of microarray data classification, the results obtained from the classification of breast cancer have the lowest accuracy, around 70%, relative to other diseases [18],[30] .

1.4 Aim and Objectives of Study

This research study aims at improving the accuracy of classification of clinical outcomes for breast cancer by gene expression profiling. To fulfillment the aim of the study, the following objectives will be achieved :

- ***(i)*** *To design and implement a gene selection technique based on integrating the ranked-partition genes with sequential selection wrapper.*
- ***(ii)*** *To obtain the optimal subset of the most informative genes, gene expression profiling, from the microarray data provided using the proposed technique.*
- ***(iii)*** *To improve the classification accuracy of the clinical outcomes of breast cancer using the obtained gene expression profiling .*

1.5 Scope of Study

For achieving the objectives of this study, a model for gene selection was proposed. This model combines the partitioned outcome of the filter technique and the wrapper technique to integrate their advantages and reduce their drawbacks. The breast cancer microarray data of Van't Veer et al. 2002 were used to investigate the proposed model of feature selection. The stages of designed technique were performed by using MATLAB as a tool where it provides wide range of powerful functions in the field of microarray data analysis. Results obtained from the implementation of the proposed technique were evaluated by comparing them with results obtained from the original studies.

1.6 Thesis Organization

This section provides general structure of this thesis, the outline is constructed as follows:

Chapter Two presents a review of the literature for previous studies related to microarray data analysis. This chapter includes a review for different techniques of machine learning that commonly applied in the context of microarray data analysis.

Chapter Three discusses a description of the methodological approach followed in this study including the data set used, the proposed model and steps of achieving the study objective .

Chapter Four explains a description for investigating the performance of the proposed feature selection model . In this chapter, implementation of steps to achieve the study objective are described. Furthermore, evaluations of results are demonstrated.

Chapter Five concludes the study by demonstrating results obtained to achieve the objectives which fulfill the aim of this study. Contribution of this study in literature as well as the desired future work are mentioned in this chapter .

REFERENCES

[1] S.Y. Kung and M.W. Mak., "Feature Selection for Genomic and Proteomic Data Mining ", in *Machine Learning in Bioinformatics*, 3rd ed. , Editor :Y.Q. Zhang and J. C. Rajapakse, New Jersey: John Wiley & Sons, Inc, 2009, Ch. 1, pp.1–46.

[2] A.T. Weeraratna and D.D. Taub , "Microarray Data Analysis: An Overview of Design, Methodology and Analysis", in *Microarray Data Analysis: Methods and Applications*, 3rd ed., Editor: M. J. Korenberg, New Jersey: Humana Press Inc., 2007, Ch.1, pp. 1–16.

[3] K. Shakya et al. , "Comparison of microarray preprocessing methods", in *Advances in Computational Biology*, Springer, 2009, Vol. 680, Ch. 16, pp. 139–147.

[4] J.P. Florida et al. , "Effect of pre-processing methods on microarray-based SVM classifiers in affymetrix genechips", in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona, Spain , pp. 1–6, 2010.

[5] Z. Wang and V. Palade , "Fuzzy Gene Mining : A Fuzzy-Based Modelfor Cancer Microarray Data Analysis", in *Machine Learning in Bioinformatics*, 3rd ed., Editor :Y.Q. Zhang and J. C. Rajapakse, New Jersey: John Wiley & Sons, Inc, 2009, Ch. 5, pp. 111–134.

[6] M. Zhiyi, C. Wensheng and S. Xueguang, "Selecting significant genes by randomization test for cancer classification using gene expression data", *Journal of Biomedical Informatics*, Vol.46 , pp. 594–601, 2013.

[7] I. Guyon et al. , " Gene selection for cancer classification using support vector machines", *Mach. Learn.* , vol. 46 , pp. 389–422 , 2002.

[8] S.A. Salem, L.B. Jack and A.K. Nandi, "Investigation of Self-Organizing Oscillator Networks for Use in Clustering Microarray Data", *IEEE Trans. Nanobioscience*, vol. 7 , pp. 65–79, 2008.

[9] S. Sebastian and F. Krzysztof ," Stable feature selection and classification algorithms for multiclass microarray data", *Biology Direc*t, Vol. 7, p. 33, October 2012.

[10] C.Sujoy and M. Anirban , "Clustering Ensemble: A Multiobjective Genetic Algorithm based Approach", *Procedia Technology*, Vol. 10, pp. 443–449, 2013.

[11] U. Marquardt,R. Galle and A.Teufel, "Molecular diagnosis and therapy of hepatocellular carcinoma (HCC): An emerging field for advanced technologies", *Journal of Hepatology*, Vol. 47, pp. 1789–1797, Aug. 2012.

[12] R. Huang et al., "A Hierarchical Method for Selecting Feature Genes from Gene-Expression Profiles", *Physics Procedia*, Vol. 33, pp. 308–314, November, 2012.

[13] M. Jamiul and R. Jianhua ,"A Steiner tree-based method for biomarker discovery and classification in breast cancer metastasis", *BMC Genomics*, Vol. 13, pp. 58–69, October 2012.

[14] M.Saghatchian et al. , "Additional prognostic value of the 70-gene signature (MammaPrint) among breast cancer patients with 4-9 positive lymph nodes", *The Breast*, Vol.22, pp. 882–690, Jan. 2013.

[15] B. Frederike et al., "Molecular imaging for monitoring treatment response in breast cancer patients", *European Journal of Pharmacology*, Vol. 717, pp. 2–11, 2013 .

[16] N. Jesmin "Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer", *Expert Systems with Applications,*Vol. 39, pp. 12371-12377, 2012 .

[17] L. Sabine et al., "Gene expression analysis in biomarker research and early drug development using function tested reverse transcription quantitative real-time PCR assays", *Methods* , Vol. 59, pp. 10–19, 2013.

[18] A. J. Basel. , "The Analysis of Microarray Data" , M.S. thesis, Dept. of Electr. Eng. and Electro. University of Liverpool, Liverpool, 2011.

[19] Xin Zhou and K. Z. Mao , "LS Bound based gene selection for DNA microarray data", *Bioinformatics*, Vol. 21, no. 8, pp. 1559–1564, Jan. 2005.

[20] S. Vanichayobon, S. Wichaidit and W. Wettayaprasit, "Microarray Gene Selection Using Self-Organizing Map", in *The 7th WSEAS International Conference on Simulation, Modelling and Optimization*, Beijing, China, pp. 239–244, September 15-17, 2007.

[21] A. J. Basel et al.,"Paradigm of Tunable Clustering Using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery", *PLoS One*, vol. 8, no. 2, Feb. 2013.

[22] A. Kulkarni et al. , "Colon cancer prediction with genetics profiles using evolutionary techniques", *Expert Syst. Appl.*, Vol. 38, Issue 3, pp. 2752–2757, Mar. 2011.

[23] W. Wang , "An Improved Non-negative Matrix Factorization Algorithm for Combining Multiple Clusterings", in *in the 2010 International Conference on Machine Vision and Human-machine Interface.*, Kaifeng , China , pp. 604–607, 24-25 April 2010.

[24] C. Xiaowei, S.Frank and Z. Hongyu, "Joint analysis of expression profiles from multiple cancers improves the identification of microRNA–gene interactions", *Bioinformatics (Gene Expression)*, Vol. 29, no. 17, pp. 2137–2145, 2013.

[25] G. Abraham et al. , "Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context", *BMC Bioinformatics*, Vol. 11, p. 277, May 2010.

[26] D. Houtao and R. George, "Gene selection with guided regularized random forest", *Pattern Recognition*, Vol. 46, pp. 3483–3489, Dec. 2013.

[27] S.Yan et al., "Identification of Novel Tissue-Specific Genes by Analysis of Microarray Databases: A Human and Mouse Model", *PLoS One*, Vol. 8, no. 5, May 2013.

[28] P. Yongjun et al. , "An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data", *Bioinformatics(Data and text mining),*Vol. 28, no. 24, pp. 3306–3315, 2012 .

[29] D. Chung and K. Roch, "Genome-Wide Analysis of Gene Expression", *Encyclopedia of Biological Chemistry*, 2nd ed., pp. 369–374, 2013 .

[30] R.S. Welsch and R.E. Menjoge , " Comparing and Visualizing Gene Selection and Classification Methods for Microarray Data", in *Machine Learning in Bioinformatics*, 3rd ed. , Editor :Y.Q. Zhang and J. C. Rajapakse, New Jersey: John Wiley & Sons, Inc, 2009, Ch. 2 , pp. 47–68.

[31] S. Pang et al. , "Bootstrapping Consistency Method for Optimal Gene Selection from Microarray Gene Expression Data for Classification Problems", in *Machine Learning in Bioinformatics*, 3rd ed. , Editor :Y.Q. Zhang and J. C. Rajapakse, New Jersey: John Wiley & Sons, Inc, 2009, Ch. 4 , pp. 89–110.

[32] G.Z. Yang and J.Y. Li , "Feature Selection for Ensemble Learning and Its Application", in *Machine Learning in Bioinformatics*, 3rd ed. , Editor :Y.Q. Zhang and J. C. Rajapakse, New Jersey: John Wiley & Sons, Inc, 2009, Ch. 6 , pp. 135–156.

[33] M. Debahut,"Discovery of Overlapping Pattern Biclusters from Gene Expression Data using Hash based PSO", *Procedia Technology*, Vol. 4,pp. 390–394, 2012.

[34] R. Giancarlo and F.Utro," Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis", *Theoretical Computer Science*, Vol. 428, pp. 58–79, 2012 .

[35] K. Mingoo Kim, C. Sung Bum and K. Ju Han Kim, "Mixture-model based estimation of gene expression variance from public database improves identification of differentially expressed genes in small sized microarray data", *Bioinformatics(Gene Expression)*, Vol. 26 no.4 , pp. 486–492, 2010.

[36] H. Michael et al.," Computational analysis of high-density peptide microarray data with application from systemic sclerosis to multiple sclerosis", *Autoimmunity Reviews*, Volume 11, , pp. 180–190, 2012.

[37] T. Ed. Kohonen, *Self-Organizing Maps*, 3rd ed. , New York, USA: Springer-Verlag, 2011.

[38] A. Neme and P. Miramontes , "A Parameter in the Learning Rule of SOM That Incorporates Activation Frequency", in *ICANN*, Athens, Greece , Vol. 4131, pp. 455–463, Sep. 10-14 2006.

[39] N.Loris ,B. Sheryl and L. Alessandra , "Combining multiple approaches for gene microarray classification", *Bioinformatics(Gene Expression)*, Vol. 28 no.8, pp. 1151–1157, 2012 .

[40] J. Nikkila et al. , "Analysis and visualisation of gene expression data using self-organizing maps", *Neural Networks*, Vol.15, pp. 953–966, Nov. 2002.

[41] M.B.H. Rhouma and H. Frigui , "Self-organization of pulse-coupled oscillators with application to clustering", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 180–195, Feb. 2001.

[42] S. Numanul et al.," Multiple gene expression profile alignment for microarray time-series data clustering", *Bioinformatics(Gene Expression)*, Vol. 26, no.18 , pp. 2281–2288, 2010.

[43] S. VEGA-PONS and JR SHULCLOPER , "A Survey of Clustering Ensemble Algorithms", *Int. J. Pattern Recognit. Artifcial Intell.*, Vol. 25, no. 3, pp. 337–372, 2011.

[44] K. Tumer and Adrian K. Agogino, "Ensemble clustering with voting active clusters", *Pattern Recognit. Lett.*, Vol. 29, no. 14, pp. 1947–1953, Oct. 2008.

[45] L. Natthakan , B. Tossapon and G. Simon," LCE: a link-based cluster ensemble method for improved gene expression data analysis",*Bioinformatics*,Vol. 26, no.12, pp. 1513-1519, 2010 .

[46] H. G. Ayad and M. S. Kamel ,"On voting-based consensus of cluster ensembles", *Pattern Recognit.*, Vol. 43 Issue 5 , pp. 1943–1953, May 2010.

[47] A. Strehl and J. Ghosh , "Cluster Ensembles – A Knowledge Reuse Modelfor Combining Multiple Partitions", *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[48] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning" ,in *The Twenty-first International Conference on Machine Learning*, New York, NY, USA , 2004.

[49] S.Vega-Pons, J.Correa-Morris and J. Ruiz-Shulcloper, "Weighted cluster ensemble using a kernel consensus function", in *Lecture Notes in Computer*

Science, J. Ruiz-Shulcloper and W.G. Kropatsch, Berlin: Springer-Verlag, 2008, Vol. 5197, pp. 195–202.

[50] V.Filkov and S. Skiena, "Integrating microarray data by consensus clustering", in *International Conference on Tools with Artificial Intelligence*, California, USA, pp. 418 – 426 , Nov. 2003 .

[51] G. Arief et al. , "Partial least squares and logistic regression random-effects estimates for gene selection in supervised classification of gene expression data", *Journal of Biomedical Informatics*, Vol. 46, pp. 697–709, 2013.

[52] C. CORTES and V. VAPNIK , "Support-Vector Networks", *Mach. Learn. J.*, Vol. 20 , Issue 3, pp. 273–297, Sep. 1995.

[53] T. Muchenxuan ," An ensemble of SVM classifiers based on gene pairs" , *Computers in Biology and Medicine*, Vol. 43, pp. 729–737, 2013.

[54] A. Filippo et al. , "Artificial neural networks in medical diagnosis", *Journal of Applied Biomedicine*, Vol. 11, pp. 47–58, 2013.

[55] H.Huey-Miin, Z. Da-Wei and T. Chen-An, "Random forests-based differential analysis of gene sets for gene expression data", *Gene*, Vol. 518(1), pp. 179–186, April 2013.

[56] M. Ioannis , "A semi-supervised fuzzy clustering algorithm applied to gene expression data", *Pattern Recognition*, Vol. 45 , pp. 637–648, 2012.

[57] A. Banumathi and A. Pethalakshmi , "Increasing cluster uniqueness in Fuzzy C-Means through affinity measure", in *Pattern Recognition, Informatics and Medical Engineering (PRIME), International Conference*, Tamilnadu, India, pp. 25–29, Mar. 2012 .

[58] P.Harun et al. , "Clustering of high throughput gene expression data", *Computers & Operations Research*,Vol.39 , pp. 3046–3061, 2012 .

[59] P. Harun et al. , "Performance of an Ensemble Clustering Algorithm on Biological Data Sets", *J. Math. Comput. Appl.*, Vol. 16, no. 1, pp. 87–96, 2011.

[60] B. Haibe-Kains, " Genome-wide gene expression profiling to predict resistance to anthracyclines in breast cancer patients", *Genomics Data,*Vol. 1, pp.7-10, 2013.

[61] J. D. Holliday et al. "Clustering Files of Chemical Structures Using the Fuzzy k-Means Clustering Method", *J. Chem. Inf. Comput.*, Vol. 44, pp. 894–902, 2004.

[62] S. Sriparna et al., "Gene expression data clustering using a multi objective symmetry based clustering technique" , *Computers in Biology and Medicine* , Vol. 43, pp. 1965–1977, 2013.

[63] A. T. Merryweather-Clarke et al. , “Global gene expression analysis of human erythroid progenitors”, *Blood*, Vol. 117, pp. 96–110, Dec. 2011.

[64] F. Rui, A.K. Nandi and Li-Yun Gong , “Clustering analysis for gene expression data: A methodological review”, in *the 5th International Symposium on Communications, Control and Signal Processing (ISCCSP)*,Rome Italy , pp. 1–6, May 2012 .

[65] Shinn-Ying Ho et al. , “Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis”, *BioSystems*, Vol. 86, no. 3, pp. 165–176, Sept. 2006.

[66] Shu-Qin Zhang et al. , “A new multiple regression approach for the construction of genetic regulatory networks”, *Artif. Intell. Med.*, Vol. 48, pp. 153–160, Mar. 2010.

[67] S. Jonnalagadda and R. Srinivasan , “ NIFTI: An evolutionary approach for finding number of clusters in microarray data”, *BMC Bioinformatics*, Vol. 10, p. 40, Jan. 2009.

[68] A. Jorge and N. Hilario, “Exploring correlations in gene expression microarray data for maximum predictive–minimum redundancy biomarker selection and classification”,*Computers in Biology and Medicine*, Vol. 43, pp. 1437–1443, 2013.

[69] F. Samah Jamal et al.,“ Complementary ensemble clustering of biomedical data”, *Journal of Biomedical Informatics*, Vol. 46, pp. 436–443, 2013 .

[70] S. Vega-Pons, J. Ruiz-Shulcloper and A. Guerra-Gandón, “Weighted association based methods for the combination of heterogeneous partitions”, *Pattern Recognit. Lett.*, Vol. 32, no. 16, pp. 2163–2170, Dec. 2011.

[71] L. J. Van't Veer et al. , “Gene expression profiling predicts clinical outcome of breast cancer”, *Nature*, Vol. 415, no. 345, pp. 530–536, Jan. 31 2002.

[72] T. R. Golub et al. , “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring ”, *Science* ., Vol. 286, pp. 531–537, Oct. 1999 .

[73] T. Hastie, R. Tibshirani and J. Friedman , " The Elements of Statistical Learning" , 2nd. ed. , Springer-Verlag , 2002 , pp.417-423,pp. 463-468 .

[74] R. Tibshirani et al. , "Diagnosis of multiple cancer types by shrunken centroids of gene expression" *PNAS*, Vol. 99, no. 10, pp. 6567-6572, May 14 2002 .

[75] C. Xiaoping et al., "Optimal combination of feature selection and classification via local hyperplane based learning strategy", *BMC Bioinformation,* Vol. 16(1), 2015.