Integrating the Principal Component Analysis with Partial Decision Tree in Microarray Gene Data

Mohammad Subhi Al-Batah

Department of Computer Science, Faculty of Science and Information Technology, Jadara University, Irbid, Jordan

Summary

In microarray cancer datasets, the gene analysis and classification is an imperative task because gene expression data have large dimensionalities, contain redundant information, irrelevant features and noises. Therefore, the main contribution of this paper is selecting a concise subset of informative genes, for improving processing speed and prediction performance. A two-phase hybrid approach is proposed which combines the Principal Component Analysis (PCA) algorithm with Partial Decision Tree (PART) rules. The PCA is applied to identify a small set with most discriminating genes, while the PART rules is proposed to classify microarray data into two or multi-classes. Eleven datasets that consists of different classes, and genes are used, which are Breast Cancer, CNS, Colon, Leukemia, Leukemia_3C, Leukemia_4C, Lung, Lymphoma, MLL, Ovarian, and SRBCT. The data analysis is conducted by using the full training method and the cross validation technique; 2-folds to 10-folds. Experimental analysis shows that gene selection using PCA method reduced the computational complexity and obtained the smallest subset of genes prior to classification. Also, it was noticed that the PART classifier when combined with PCA algorithm works faster and showed a remarkable improvement in the classification accuracy.

Key words:

Principal Component Analysis (PCA) algorithm, Partial Decision Tree (PART) rules, Microarray data, Classification, Gene selection, Data mining

1. Introduction

Cancer is among the leading causes of death worldwide, thus prediction and classification of cancer types is a first order task in the medical sector [1-3]. Microarray data usually contains redundant and irrelevant features (genes). These features increase the computational burden and negatively affect the performance of the classifier [4]. Hence, it is desirable to perform feature selection to detect and select relevant, non-redundant and interacting genes in an efficient way [5-7]. Feature selection is a preprocessing phase which aims to improve the accuracy, speed, data quality, and data understanding. It also serves to reduce dimensionality and computational resources [8].

Two main techniques are used in feature selection which include wrapper and filter methods. Wrapper model approach uses the method of classification itself to measure the importance of features set, hence the feature selected depends on the classifier model used [9]. The filter approach actually precedes the actual classification process. The filter approach is independent of the learning algorithm, computationally simple, fast and scalable. Feature selection using filter method is done once and then can be provided as input to different classifiers [10].

Various feature ranking and feature selection techniques have been adopted in the literature such as Principal Component Analysis (PCA), Information Gain (IG), Gain Ratio (GR), Symmetric Uncertainty (SU), Mutual Information (MI), Gini Index (GI), Chi-Square, Euclidean Distance, T-test, minimum Redundancy and maximum Relevance (mRmR), Fisher score, Pearson Correlation Coefficient, Crammer's V, Markov's Blanket Filter (MBF), Random Forest, Kruskal Wallis, Laplacian Score, SPEC, Correlation-based Feature Selection (CFS), Fast Correlation Based Filter (FCBF), Relief, Relief-F, Las Vegas Filter (LVF), FOCUS, One-R, Kolmogorov-Smirnov Feature Filter, Pearson’s Redundancy Based Filter (PRBF), INTERACT, Feature Selection Based on Mutual Correlation, Incremental Usefulness, CorrSF and ConsSF [11].

Some of these filter methods do not perform feature selection but only feature ranking hence they are combined with a search method when one needs to find out the appropriate number of attributes [12]. Such filters are often used with forward selection, which considers only additions to the feature subset, backward elimination, bi-directional search, best-first search, genetic search, greedy stepwise, ranker search, and other methods [13]. In this paper, Ranker search is used as a search method with Principal Component Analysis (PCA) as a subset evaluating mechanism.

Classification is the technique to categorize the data into a desired and distinct number of classes according to particular characteristics [14-16]. Approaches based on machine learning, which can automatically acquire qualitatively interesting patterns from gene data, have been widely adopted. Among these machine learning approaches used to study performance of microarray data are: support vector machine (SVM) [17], artificial neural network (ANN) [18] and fuzzy decision tree algorithm [19].

The authors Hala et al. [20] adopted a hybrid gene selection namely Genetic Bee Colony (GBC) in microarray dataset. In GBC, both Genetic Algorithm and

Manuscript received March 5, 2019
Manuscript revised March 20, 2019

Artificial Bee Colony have been applied to select the most informative and predictive genes for microarray classification. Zheng et al. [21] used independent component analysis to refine a subset of genes to further improve the clustering performance of nonnegative matrix factorization. Yan et al. [22] applied the sparse representation-based classification (SRC) scheme in the diagnosis of microarray gene expression in cancer. The SRC showed better performance than the state-of-the art methods. Zhu et al. [23] applied the Markov Blanket-Embedded Genetic Algorithm (MBEGA) for gene selection problem. The embedded Markov blanket-based memetic operators add or delete features (genes) from a Genetic Algorithm (GA) solution so as to quickly improve the solution and fine-tune the search. Kar et al. [24] applied a PSO–adaptive K-nearest neighbor (KNN) based gene selection method and they used a heuristic for selecting the optimal values of K, while the classification accuracies have been tested using SVM algorithm. Hameed et al. [25] used a hybrid method which combines three filters with geometric binary particle PSO and SVM for effective gene selection and classification in the high dimensional data.

In this paper, a two-phase hybrid form of PCA and PART is proposed to perform effective selection and classification task in the high dimensional microarray data. The PCA is applied to select relevant features. Then, the PART rules is applied to classify microarray dataset into cancerous/non-cancerous. The proposed method is applied to eleven microarray datasets which include Breast Cancer, CNS, Colon, Leukemia, Leukemia_3C, Leukemia_4C, Lung, Lymphoma, MLL, Ovarian, and SRBCT. To find the best performance, the full training and cross validation techniques are used in the analysis [26-28].

2. Theoretical Consideration

2.1 Principal Component Analysis (PCA)

Feature selection help to improve the performance of learning models by alleviating the effect of the curse of dimensionality, enhancing generalization capability, and speeding up learning process. Also, feature selection helps researchers to acquire better understanding about the data [29]. PCA is a popular linear feature extractor used for unsupervised feature selection based on eigenvectors analysis to identify the critical original features for a principal component [30].

PCA is a useful linear transformation technique that is used in numerous applications, such as face recognition and image compression, stock market predictions, analysis of gene expression data, and many more [31]. The goal of PCA [32] is to find a set of new attributes (PCs) which meets the following criteria: The PCs are (i) linear combinations of the original attributes, (ii) orthogonal to each other, and (iii) capture the maximum amount of variation in the data. The variability of the data can be captured by a relatively small number of PCs, and, as a result, PCA can achieve high dimensionality reduction with lower noise than the original patterns. In this paper, principal component's algorithm is used in conjunction with a Ranker search. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for 95% of the variance in the original data.

2.2 Partial Decision Tree (PART)

PART is a partial decision tree algorithm, which is a combination of C4.5 and RIPPER rule learning. PART is a separate-and-conquer rule learner proposed by Witten and Eibe [33]. The algorithm produces sets of rules called decision lists which are pre-ordered. New data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule [34-36].

2.3 Microarray Datasets

In this paper, the datasets represent eleven high dimensional microarray datasets for different types of disease [23]. The datasets include Breast Cancer, CNS, Colon, Leukemia, Leukemia_3C, Leukemia_4C, Lung, Lymphoma, MLL, Ovarian, and SRBCT. The main characteristics of the datasets such as the number of total genes, the number of instances, and the number of classes are summarized in Table 1. As it can be seen in table 1, the number of genes (features) is so high, whereas the number of instances (samples) is so low in all datasets. This is exactly the challenge when microarray data are involved. For example, the “Colon Tumor” dataset contains only 62 samples with 2000 genes. Thus, classification methods cannot perform well because of the “curse of dimensionality” phenomena, where excessive features may actually degrade the performance of a classifier if the number of training examples used to build the classifier is relatively small compared to the number of features [37].

Table 1: Summary of gene microarray datasets.

Dataset	#Gene	#Instance	#Class
Breast Cancer	24481	97	2 classes 46 relapse, 51 non-relapse
Central Nervous System	7129	60	2 types 21 survivors, 39 failures
Colon Tumor	2000	62	2 types 40 Tumor, 22 Normal
Leukemia	7129	72	2 types of acute leukemia 47 Acute Lymphoblastic Leukemia (ALL), 25 Acute Myeloid Leukemia (AML)
Leukemia_3C	7129	72	3 types of acute leukemia 38 B-cell ALL, 9 T-cell ALL, 25 AML
Leukemia_4C	7129	72	4 types of acute leukemia 38 B-cell, 9 T-cell, 21 BM AML, 4 PB AML
Lung Cancer	12600	203	5 types 139 adenocarcinoma (AD), 17 normal lung (NL), 6 small cell lung cancer (SMCL), 21 squamous cell carcinoma (SQ), 20 pulmonary carcinoid (COID).
Lymphoma	4026	66	3 different adult lymphoid malignancies 46 diffuse large B-cell lymphoma (DLBCL), 9 Follicular Lymphoma (FL), 11 Chronic Lymphocytic Leukemia (CLL).
Mixed Lineage Leukemia (MLL)	12582	72	3 types 24 acute lymphoblastic leukemia (ALL), 20 Mixed-Lineage Leukemia (MLL), 28 acute myeloblastic leukemia (AML).
Ovarian Cancer	15154	253	2 types 162 Cancer, 91 Normal
Small Round Blue-Cell Tumor (SRBCT)	2308	83	4 different cases 29 Ewing sarcoma (EWS), 11 Burkitt lymphoma (BL), 18 neuroblastoma (NB), 25 rhabdomyosarcoma (RMS).

3. Experimental Consideration

3.1 Features Selection Experiment

In this paper, 11 different high dimensional datasets are applied to test the applicability of the proposed method. In feature selection stage, the PCA with Ranker search method and full training data is conducted for each individual microarray gene dataset. The PCA is considered in order to reduce the computational complexity and remove the irrelevant gens. After applying PCA on datasets, we have a new subset with a small size of dimension. Table 2 shows the number of selected gens, which have strong discriminating capacity to distinguish the samples into different classes. From the obtained result, it is inferred that, the number of selected genes by PCA is so lower than the number of initial genes. For example, in Leukemia, total features are equal to (7129) while we have (60) selected features using PCA. The results also prove that the PCA is able to decrease the data size and then reduces the time taken by the classifier to complete the classification job.

Table 2: Number of selected genes before/after applying PCA algorithm

Dataset	# Total Genes	# Gene After PCA
Breast Cancer	24481	54
CNS	7129	45
Colon Tumor	2000	31
Leukaemia	7129	60
Leukaemia-3C	7129	60
Leukaemia-4C	7129	60
Lung Cancer	12600	63
Lymphoma	4026	51
MLL	12582	58
Ovarian Cancer	15154	42
SRBCT	2308	61

3.2 Classification Experiment

In the classification stage, the PART rules is applied on the original datasets. After that, the PART rules is applied on each newly obtained dataset containing only the selected genes. To evaluate the genes classification accuracy, the full training data and the cross validation are utilized. For the cross validation, the methods used are 2-fold to 10-fold.

Table 3: The PART classification results before applying PCA

Test Method	Breast Cancer	CNS	Colon	Leukemia	Leukemia_3C	Leukemia_4C	Lung	Lymphoma	MLL	Ovarian	SRBCT
Full Training	97.94	98.33	98.39	98.61	98.61	98.61	99.01	100.00	98.61	100.00	98.80
2-Fold	59.79	60.00	83.87	76.39	73.61	62.50	80.79	84.85	81.94	95.65	85.54
3-Fold	57.73	60.00	79.03	87.50	81.94	77.78	89.16	87.88	88.89	95.65	78.31
4-Fold	52.58	55.00	82.26	86.11	79.17	77.78	88.18	89.39	83.33	97.63	73.49
5-Fold	56.70	55.00	83.87	84.72	93.06	79.17	88.67	89.39	83.33	98.42	78.31
6-Fold	55.67	55.00	79.03	83.33	83.33	86.11	91.13	89.39	90.28	97.63	80.72
7-Fold	62.89	55.00	75.81	76.39	84.72	75.00	89.66	96.97	87.50	98.02	74.70
8-Fold	60.82	55.00	74.19	79.17	91.67	79.17	89.66	93.94	88.89	98.42	75.90
9-Fold	55.67	61.67	82.26	81.94	87.50	79.17	86.21	93.94	80.56	98.42	79.52
10-Fold	62.89	56.67	82.26	83.33	93.06	83.33	91.13	92.42	86.11	97.63	79.52
Average	62.27	61.17	82.10	83.75	86.67	79.86	89.36	91.82	86.94	97.75	80.48

Table 4: The PART classification results after applying PCA

Test Method	Breast Cancer	CNS	Colon	Leukemia	Leukemia_3C	Leukemia_4C	Lung	Lymphoma	MLL	Ovarian	SRBCT
Full Training	97.94	98.33	98.39	98.61	98.61	98.61	99.51	100.00	98.61	100.00	98.80
2-Fold	63.92	56.67	77.42	77.78	91.67	73.61	86.21	87.88	87.50	96.05	85.54
3-Fold	63.92	68.33	83.87	87.50	81.94	81.94	91.63	90.91	83.33	95.65	79.52
4-Fold	74.23	70.00	87.10	90.28	91.67	91.67	87.68	96.97	86.11	98.42	83.13
5-Fold	76.29	76.67	82.26	88.89	93.06	83.33	91.63	90.91	87.50	98.42	84.34
6-Fold	63.92	71.67	90.32	80.56	84.72	84.72	88.18	96.97	91.67	97.63	83.13
7-Fold	72.16	71.67	85.48	80.56	84.72	84.72	91.63	96.97	87.50	98.81	79.52
8-Fold	68.04	68.33	88.71	84.72	93.06	88.89	91.13	95.45	93.06	99.21	84.34
9-Fold	78.35	73.33	88.71	86.11	87.50	83.33	91.13	96.97	91.67	99.21	83.13
10-Fold	71.13	65.00	87.10	84.72	93.06	88.89	91.13	95.45	90.28	98.81	83.13
Average	72.99	72.00	86.94	85.97	90.00	85.97	90.99	94.85	89.72	98.22	84.46

Table 3 and Table 4 summarize the accuracy of PART on 11 Microarray datasets after and before applying PCA using full training and cross validation methods. The results show that generally the accuracy of the PART on the filtered dataset performed better results when compared with those applied directly on the original datasets. However, there are some datasets in which the accuracy on the original dataset is same as the filtered dataset. From the results, the average accuracy of the PART when using PCA as compared to the PART with original datasets is increased over 10.72% for Breast Cancer, 10.83% for CNS, 4.84% for Colon, 2.22% for Leukemia, 3.33% for Leukemia_3C, 6.11% for Leukemia_4C, 1.63% for Lung, 3.03% for Lymphoma, 2.78% for MLL, 0.47% for Ovarian, and 3.98% for SRBCT.

In addition, it can be seen in Table 3, and 4 that the full training method has presented the highest classification accuracy as compared to cross validation method. For example, on Lung cancer, the accuracy of the PART on original datasets are 99.01 (full Training), 80.79 (2-fold), 89.16 (3-fold), 88.18 (4-fold), 88.67 (5-fold), 91.13 (6-fold), 89.66 (7-fold), 89.66 (8-fold), 86.21 (9-fold), and 91.13 (10-fold). Also, the accuracy of the PART after PCA is 99.51, 86.21, 91.63, 87.68, 91.63, 88.18, 91.63, 91.13, 91.13, and 91.13 for full Training, 2-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, and 10-fold, respectively.

Considering the 10-fold cross validation method. The comparison between the accuracy by PART before and after PCA using 10-fold cross validation is given in Fig. 1. What is clear in Fig. 1, the accuracy has increased when PART is applied on the selected genes which were obtained after applying PCA as compared on the original data. This indicates that the feature selection by PCA not only improved the efficiency of the classification process but also its accuracy was enhanced.

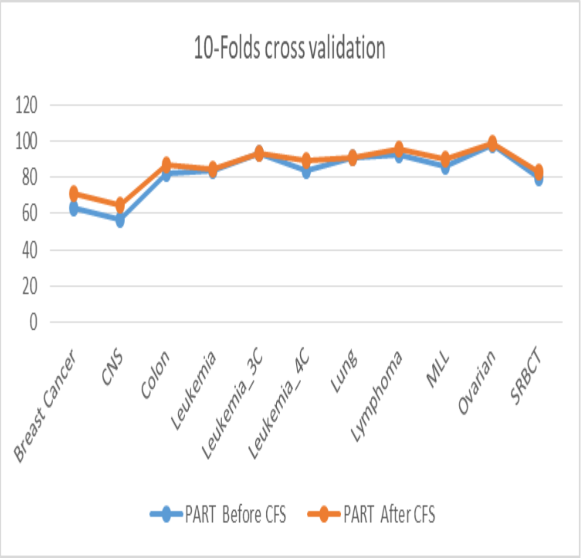

Fig. 1 Accuracy of PART before/after PCA using 10-fold cross validation

4. Conclusion

The gene microarray selection and classification is considered challenging problem for the diagnosis of disease and cancers. In this paper, the proposed method is composed of two-phase hybrid form of PCA algorithm and PART rules. The experiment is conducted with 11 datasets and the result proved that the proposed method is efficient for selecting effective genes and enhancing predictive accuracy. In most cases, the best accuracy is achieved when we applied the full training method as compared to the cross validation techniques. In addition, the outcome shows that the proposed method is powerful, stable, less complex, and suitable for gene microarray classification. In the future, we plan for using different percentages of distribution for training and testing datasets, and considering the applicability of another machine learning techniques such as Genetic Algorithm, Neural Networks, and Fuzzy Logic, etc.

Acknowledgment

This research was supported by Jadara University, Jordan. The authors would like to express their cordial thanks to Dr. Malak Fuad Subhi Al-Battah, a Neurology resident in King Abdullah University Hospital (KAUH), for her valuable advice about microarray cancer datasets and comments that greatly improved the study and provided insight and expertise that greatly assisted the research.

References

[1] N. R. Mat Noor, N. A. Mat Isa, and M. S. Al-Batah, “Automatic glass-slide capturing system for cervical cancer pre-screening program,” American Journal of Applied Sciences, vol.5, no.5, pp.461-467, 2008.

[2] A. Quteishat, M. al-batah, A. al-mofleh, and S. H. Alnabelsi, “Cervical Cancer Diagnostic System Using Adaptive Fuzzy Moving K-means Algorithm and Fuzzy MIN-MAX Neural Network,” Journal of Theoretical and Applied Information Technology, vol.57, no.1, pp.48-53, 2013.

[3] M. S. Al-batah, “Testing the Probability of Heart Disease Using Classification and Regression Tree Model,” Annual Research & Review in Biology, vol.4, no.11, pp.1713-1725, 2014.

[4] L. Wang, 2012, “Feature selection in bioinformatics,” Proceedings of the Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems and Nanoengineering X. May 10, SPIE. DOI: 10.1117/12.921417

[5] M. Sh. Alkhasawneh, U. Ngah, L. T. Tay, M. S. Al-batah, and N. A. Mat Isa, “Intelligent Landslide System Based on Discriminant Analysis and Cascade-Forward Back-Propagation Network,” Arabian Journal for Science and Engineering, ISSN: 1319-8025, Springer, HEIDELBERG, GERMANY, 2014.

[6] M. S. Al-Batah, A. Zabian, and M. Abdel-wahed, “Suitable Features Selection for the HMLP Network using Circle Segments Method,” European Journal of Scientific Research, vol.67, no.1, pp.52-65, 2011.

[7] M. Sh. Alkhasawneh, U. Ngah, L. T. Tay, M. S. Al-batah, and N. A. Mat Isa, “Determination of Important Topographic Factors for Landslide Mapping Analysis Using MLP Network,” Scientific World Journal, ISSN: 1537-744X, vol.2013, Article ID 415023, 12 pages.

[8] R.K. Singh, and M. Sivabalakrishnan, “Feature selection of gene expression data for cancer classification: A review,” Procedia Computer Sci., vol.50, pp.52-57, 2015.

[9] V. Bolón-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos, “A review of feature selection methods on synthetic data,” Knowl. Inform. Syst., vol.34, pp.483-519, 2013.

[10] Z. M. Hira, and D. F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data,” Advances Bioinformatics. DOI: 10.1155/2015/198363, 2015

[11] K. Mani, and P. Kalpana, “A Review on Filter Based Feature Selection,” International Journal of Innovative Research in Computer and Communication Engineering, vol.4, no.5, pp.9146-9156, 2016.

[12] M. S. Al-Batah, “Automatic Diagnosis System for Heart Disorder using ESG Peak Recognition with Ranked Features Selection,” international journal of circuits, systems and signal processing, ISSN: 1998-4464, vol.13, pp.391-398, 2019.

[13] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection,” International Journal of Information Technology and Knowledge Management, vol.2, no.2, pp.271-277, 2010.

[14] A. K. Baareh, A. F. Sheta, and M. S. Al-Batah, “Feature based 3D Object Recognition using Artificial Neural

Networks," International Journal of Computer Applications, vol.44, no.5, pp.1-7, 2012.

[15] N. A. Mat Isa, Z. M. Sani, and M. S. Al-Batah, "Automated Intelligent real-time system for aggregate classification," International Journal of Mineral Processing, vol.100, no.1-2, pp.41-50, 2011.

[16] M. S. Al-batah, M. Sh. Alkhasawneh, L. T.Tay, U. Ngah, H. H. Lateh, and N. A. Mat Isa, "Landslide Occurrence Prediction Using Trainable Cascade Forward Network and Multilayer Perceptron," Mathematical Problems in Engineering, vol. 2015, Article ID 512158, 9 pages.

[17] S. Cogill, and L. Wang, "Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates," Bioinformatics, vol.32, no.23, pp.3611-3618. PMID: 27506227, 2016.

[18] R. Aziz, C. K. Verma, M. Jha and N. Srivastava, "Artificial neural network classification of microarray data using new hybrid gene selection method," Int. J. Data Mining Bioinform, vol.17, pp.42-65, 2017.

[19] S. A. Ludwig, S. Picek, and D. Jakobovic, "Classification of Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm," In: Operations Research Applications in Health Care Management, C. Kahraman and Y.I. Topcu (Eds.), Cham: Springer International Publishing, pp.327-347, 2018.

[20] M. A. Hala, H. B. Ghada, and A. A. Yousef, "Genetic Bee Colony (GBC) Algorithm: A new Gene Selection Method for Microarray Cancer Classification," Computational Biology and Chemistry, vol.56, pp.49-60, 2015.

[21] C. H. Zheng, D. S. Huang, L. Zhang, and X. Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," IEEE Transactions on Information Technology in Biomedicine, vol.13, no.4, pp.599-607, 2009.

[22] K. Yan, Y. Xu, X. Fang, C. Zheng, and B. Liu, "Protein fold recognition based on sparse representation based classification," Artificial Intelligence in Medicine. https://doi.org/10.1016/j.artmed.2017.03.006 PMID: 28359635, 2017.

[23] Z. Zhu, Y. S. Ong, and M. Dash, "Markov Blanket-Embedded Genetic Algorithm for Gene Selection," Pattern Recognition, vol.49, no.11, pp.3236-3248, 2007.

[24] S. Kar, K. Das Sharma, and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique," Expert Systems Applications, vol.42, pp.612-627, 2015.

[25] S. S. Hameed, R. Hassan, and F. F. Muhammad, "Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm," PLOS ONE, 12: e0187371, 2017.

[26] N. A. Mat-Isa, A. Joret, M. S. Al-Batah, A. N. Ali, K. Z. Zamli, and K. A. Azizli, "Microcontroller Based HMLP Realization for Aggregate Classification System," International Journal of Factory Automation, Robotics and Soft Computing, no.2, pp.19-26, 2006.

[27] Z. Md-Sani, N. A. Mat-Isa, S. A. Suandi, M. S. Al-Batah, K. Z. Zamli, and K. A. Azizli, "Intelligent Rock Vertical Shaft Impact Crusher Local Database System," International Journal of Computer Science and Network Security, vol.7, no.6, pp.57-62, 2007

[28] M. S. Al-Batah, S. Mrayyen, and M. Alzaqebah, "Arabic Sentiment Classification using MLP Network Hybrid with Naive Bayes Algorithm," Journal of Computer Science, vol.14, no.8, pp.1104-1114, 2018,

[29] V. Singh, and S. Pathak, "Feature Selection Using Classifier in High Dimensional Data," CoRR, abs/1401.0898, 2014.

[30] F. Song, and G. Zhongwei, and M. Dayong, "Feature Selection Using Principal Component Analysis," vol.14, pp.27-30, 2010.

[31] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," Neuroinformatics, vol.12, no.2, pp.229-244, 2014.

[32] I. T. Jolliffe, and J. Cadima, "Principal component analysis: a review and recent developments," Philos Trans A Math Phys Eng Sci. 2016.

[33] I. H. Witten, and F. Eibe, "Data mining: practical machine learning tools and techniques," 2nd ed, pp.1-525, 2005.

[34] S. Swami, and O. Jangir, "A Review Paper on Data Mining Techniques and Algorithms," International Journal on Recent and Innovation Trends in Computing and Communication, vol.5, no.3, pp. 536-539, 2017.

[35] S. Mrayyen, M. S. Al-Batah, and M. Alzaqebah, "Investigation of Naive Bayes Combined with Multilayered Perceptron for Arabic Sentiment Analysis and Opinion Mining," International Journal of Mathematical Models And Methods in Applied Sciences, vol.12, 2018.

[36] M. F. J. Klaib, M. S. Al-batah, and R. J. Rasrasc, "3-way Interaction Testing using the Tree Strategy," Journal of Procedia Computer Science, International Conference on Communication, Management and Information Technology, vol.65 , pp.845–852, 2015.

[37] P. Duda, D. G. Stork, "Pattern Classification," Wiley-Interscience Publication: Hoboken, NJ, USA, 2001.

Mohammad Subhi Al-Batah received his PhD in Computer Science/ Artificial Intelligence from the University of Science Malaysia in 2009. After working as an assistant professor (from 2009) in the Dept. of Computer Science, Jadara Univ. in Jordan, he has been an associate professor at Jadara Univ. since 2014. He worked as a dean of Faculty of Science and Information Technology from 2015-2018. In 2019, He is the director of the center for Academic Development and Quality Assurance. His research interests include Image Processing, Artificial Intelligence, Medical Analysis, Real Time Classification and Software Engineering, E-mail: albatah@jadara.edu.jo, dralbatah@gmail.com.