Classification Models Based-on Incremental Learning Algorithm and Feature Selection on Gene Expression Data

Phayung Meesad[1], **Sageemas Na Wichian**[2],
Unger Herwig[3], and **Patharawut Saengsiri**[4], Non-members

ABSTRACT

This Gene expression data illustrates levels of genes that DNA encodes into the protein such as muscle or brain cells. However, some abnormal cells may evolve from unnatural expression levels. Therefore, finding a subset of informative gene would be beneficial to biologists because it can identify discriminative genes. Unfortunately, genes grow up rapidly into the tens of thousands gene which make it difficult for classifying processes such as curse of dimensionality and misclassification problems. This paper proposed classification model based-on incremental learning algorithm and feature selection on gene expression data. Three feature selection methods: Correlation based Feature Selection (Cfs), Gain Ratio (GR), and Information Gain (Info) combined with Incremental Learning Algorithm based-on Mahalanobis Distance (ILM). Result of the experiment represented proposed models CfsILM, GRILM and InfoILM not only to reduce many dimensions from 2001, 7130 and 4026 into 26, 135, and 135 that save time-resource but also to improve accuracy rate 64.52%, 34.29%, and 8.33% into 90%, 97.14%, and 83.33% respectively. Particularly, CfsILM is more outstanding than other models on three public gene expression datasets.

Keywords: Feature Selection, Incremental Learning Algorithm, Gene Expression, Classification

1. INTRODUCTION

The Central Dogma of Molecular Biology creates gene expression data which Deoxyribonucleic Acid or DNA translates into protein. However, some abnormal cells may evolve from unnatural expression levels. Nowadays, microarray techniques are used as simple method to evaluate gene expression level. This method is applied to hybridization of nucleic acid which measures tens of thousands gene expressions at the same time [1]. Nevertheless, the number of genes grows rapidly which leads to curse of dimensionality and miss classification. Thus, biologists need more time to calculate and search discriminate genes.

Motivation of incremental learning concept comes from an idea of shape increasing with new observation. Thus, keeping continuous data is the basic key to respond with new driving force [2]. In contrast, the objective of conceptual learning is to separate into training and testing sets and then the training set is learned by the numbers of target class. After that the model is created using training processes. Then the testing set is used to evaluate this model. A drawback of this technique is not having a process for new learning, for example, the stream of input data from the internet [2],[3]. Since the numbers of gene are based on microarray experiment, and the data is very large. Therefore, incremental learning procedure is suitable for creating classification model. Nevertheless, most of the genes are not related to diseases. Hence, feature selection techniques can recognize a gene that has been discriminated power. Here, it is not only finding a subset of gene but also decreasing time consumption.

A basic method for gene selection is searching the subset of genes that has more ability of classification. For now, feature selection has two techniques which are filter and wrapper approaches, first of all they consider the discriminate power of a gene one by one but do not consider the induction of an algorithm. The second technique is associated with an induction algorithm that determines the accuracy of a chosen sub group of genes. For time consumption, the wrapper approach consumes longer than the filter method. Although the correctness of the filter method is lower than the wrapper approach. In addition, feature transformation such as Single Value Decomposition (SVD), Independent Component Analysis (ICA), and Principle Component Analysis (PCA) are not appropriate for gene selecting because they cannot reduce any dimension, safeguard unrelated feature and hard

Manuscript received on July 31, 2011 ; revised on December 1, 2011.

[1] The author is with Department of Information Technology Faculty of Information Technology, KMUTNB, Bangkok Thailand. , E-mail: pym@kmutnb.ac.th

[2] The author is with Department of Applied Science and Social College of Industrial Technology, KMUTNB, Bangkok, Thailand. , E-mail: sgm@kmutnb.ac.th

[3] The author is with Department of Communication Network, Faculty of Mathematics and Computer Science, Fern University in Hagen, Germany., E-mail: Herwig.Unger@gmx.de

[4] The author is with Division of Information and Communication Technology, Thailand Institute of Scientific and Technological Research, Pathum Thani, Thailand. , E-mail: pvs@tistr.or.th

to interpret significant of genes [4].

Referring to the above; gene expression data contains large features meanwhile most algorithms have low efficiency when working with high dimensional data. Hence, this paper proposed incremental learning algorithm combining a feature selection method on gene expression data which provides solution for developing biological dimension reduction application. The result of the experiment depicted an accomplishment of these classification methods, it can decrease many dimensions and improve accuracy correctly. This paper is organized as follows; Section 2 represents a summary of the literature review. In Section 3, the proposed method is explained in detail. The result of experiment will be represented in Section 4. Finally, conclusion will be shown in Section 5. The problem of data analysis based on high dimension can be found in many domains such as document clustering multimedia, and molecular biology data. This trouble is well-known about very sensitive with curse of dimensionality. In this way, if several attributes or variables are not in the same direction, the data will be sparse that is some spaces may not exist at all. These properties can lead to be 1) expensive cost such as memory and storage usage 2) difficult for understanding the data for instance, data containing low or high dimension can be interpreted as a same group and other. Therefore, feature selection techniques are more beneficial for conducting good feature subset that can make classification algorithm effectively and efficiently. The objectives of feature selection are following [5]:

1. Reducing feature set: predicting and forecasting result based on searching expected knowledge.
2. Decrease learning time: several algorithms contained many attributes that spent more time for classification.
3. Increasing accuracy rate: noise and irrelevance features should be eliminated according to they are correlated with prediction efficiency.
4. The average values of features between low and high are the same.

2. LITERATURE REVIEWS

2.1 Feature Selection (FS)

The key objectives of FS are selecting the best position and decreasing feature more than traditional dataset. This method is different from feature extraction and transformation algorithm because it is not encoded new feature subset into another form. In contrast, feature extraction and transformation techniques make new feature subset in a new format and create a drawback of physical translating in specialist area. FS method can fall into supervised and unsupervised learning that shown in Fig. 1.

In general, feature selection in supervised learning cooperates with searching method for discovering feature subsets space. On the one hand, the objective of unsupervised feature selection is not clear two cases of unsupervised problem are prior knowledge about a many clusters and complexity of finding subset of each feature. Nevertheless, this paper focuses on only supervised learning feature selection approach.

2.1.2 Filtering Approach

As mention before, filtering approach tries to remove irrelevant feature from original feature set then send to learning algorithm. Typically, traditional dataset is analyzed for identifying influence dimension subset used to describe and relate with their structure. Hence, the filtering process is not depended on efficiency of classification algorithm for instance, Correlation Based Feature Selection (Cfs), Information Gain (IG), Chi-Square, and Gain Ratio (GR).

2.1.3 Wrapper Approachh

In other word, relevance feature subset is chosen using classifier algorithm power. In that case, searching method of wrapper based on feature space and evaluated subset value using correct estimation of predicted classification. The goal of method is finding feature subset that is proximate with criterion, for example Forward Selection (FS), Backward Elimination, Genetic Search, and Simulated Annealing.

2.1.4. Embedded Approach

Integrating between classification algorithm and feature selection technique is applied together for Embedded Approach. Actually, this technique is working with classifying and searching feature at the same time. For example, Random Forest technique based on decision tree and forced for giving score into crucial feature.

However, combining filter and wrapper approach is very interesting in gene expression data domain for example comparison of hybrid feature selection models is proposed by [6]. This research consists of four steps 1) ranking feature subset using filter approach 2) sending the result of first step into wrapper approach based on Support Vector Machine (SVM) classifier algorithm. 3) creating hybrid feature selection model such as CFSSVMGA and GRSVMGS. 4) comparing accuracy rate during hybrid feature selection models.

In the first place, feature selection technique of gene expression data is performed based on ranking systems such as difference of means and t-statistics. T-statistics has higher efficiency, than difference of mean which variance is assumed to be equal. Significant Analysis of Microarray (SAM) is proposed by [7]. SAM is complete exceeding a range of restriction that affirms genes. SAM produces individual scores of gene based on cooperation of gene expression that is associated with standard deviation of evaluation repeating.

Currently, feature selection method focus on filter and wrapper approaches [8]. In filter approach, Correlation technique is very simple and wildly used in gene selection domain. For instance, if correla-

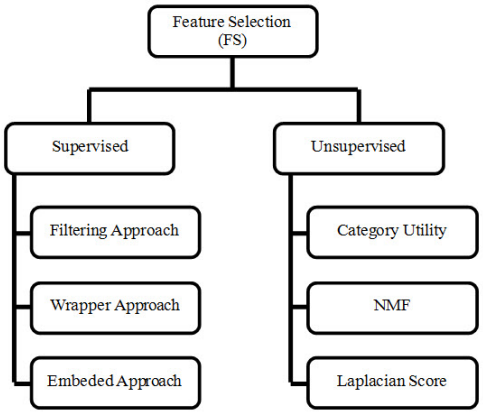

Fig.1: *Supervised and Unsupervised Learning of Feature Selection.*

tion value of each gene is higher than a threshold, these genes are integrated together. After that, gene ranking is created using correlation value and then selecting top-rank gene from gene ranking. In contrast, gene has correlation value less than threshold that depicts the best gene. Correlation method combined with Fuzzy Clustering for classification is proposed in [9]. However, disadvantage is how to define a number of clusters. Another feature selection based on gene expression data was proposed by [8] composed of two phases. 1) Cfs algorithm is used for selecting gene. 2) Binary Particle Swarm Optimization (BPSO) is used to select feature from previous phase. Nevertheless, parameter setting is very difficult for general user.

Gene boosting technique was proposed by [10] which joins the filter and wrapper approaches. Firstly, filter approach selects a subset of gene from the top rank. Next, wrapper approach chooses the output of the first step. The method will stop when accuracy rate of training set is acceptable or completed iteration. On the contrary, the overfitting problem may happen when this technique is boosted. Gene expression data contains a large feature and small instances. Thus, [11] proposed integrating two techniques, BPSO and K-nearest neighbour (K-NN) and then measured by Leave-one-out cross-validation (LOOCV). Nonetheless, the primary problem of BPSO is making local minimum from large dimension of genes.

2.2 Classification Model for Gene Expression Data

Fuzzy clustering by Local Approximation of Membership (FLAME) has different weight structure used to assign membership degree of gene for clustering. Actually, membership function of Fuzzy C-Mean (FCM) is used to weight each gene for clustering which is the factor similarity of gene comparing with mean value of cluster. FLAME algorithm, gene is specified suitable membership degree using their neighbour membership and repeating this process for converging [12]. However, it is very sensitive to adjust parameter and order of input. Incremental filter and wrapper approach combining is proposed by [13], consist of two steps. 1) Gene is evaluated using Info method and then 2) Rank list is created using wrapper method. Three algorithms are used to process: Naïve Bayes, Instance-based Learner (IB1), and Decision Tree (C4.5). Discriminate gene is determined using high score value of gene expression level. Classification of microarrays to nearest centroid (CalNC) is proposed by [14] based on Linear Discriminant Analysis (LDA) and t-statistics value. However, CalNC is updated by [15] using suitable scores. Distance during each object and centroid are evaluated using approximate score. A gene is nearly centroid is chosen using Mahalanobis distance. Multi-objective strategy based on genetic algorithm to select gene is proposed by [16]. Since one-objective of GASVM is not suitable for performing gene that is less than 1,000. Therefore, the concept of multi-objective optimization (MOO) based on relevance between many objectives and classes are replaced which is called MO-GASVM.

Conducting high accuracy rate from classification algorithm is possible in several datasets but three examples below represent that may be disadvantageous determination when ignore joining with other feature selection techniques. Because of general gene expression dataset contains many irrelevant attributes that led to miss-classification algorithm. For example, original dimension of three gene expression datasets 1) Colon Tumor 2) Leukemia and 3) Lymphoma represented low accuracy rate using ILM algorithm 64.52%, 34.29%, and 8.33%, respectively (Fig. 2).

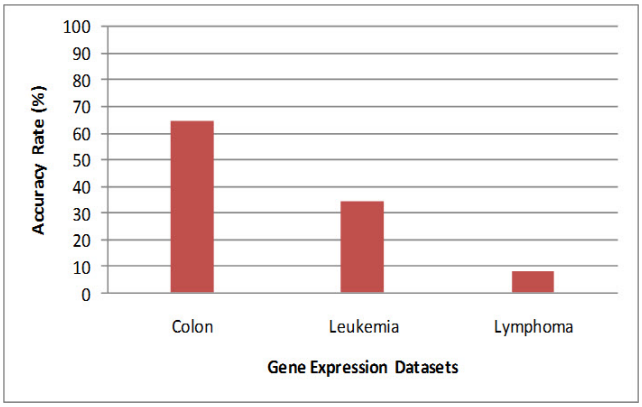

Fig.2: *The Accuracy Rate of Traditional ILM Alogrithm without Feature Selection Technique.*

3. METHODS

3.1 Correlation Based Feature Selection (Cfs)

The concept of Cfs technique is relevance of feature and target class based on heuristic operation [17]. This method was created by Mark in 1999. However, evaluation processing of Cfs technique is assigning large relation of feature subset with target class and ignores correlation. Cfs equation is that represents in (1).

$$M_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \tag{1}$$

where M_s is heuristic "merit" of feature subsets; S is a set including k features; $\bar{r}_{cf}$ is average feature-class relation ($f \in S$); and $\bar{r}_{ff}$ is mean feature-feature inter-relation.

3.2 Information Gain (Info)

An info algorithm is the most popular technique to find the node impurity; this is based on the crucial concept for splitting power of a gene [18]. An Info algorithm decides that feature via Entropy estimation. Entropy at a given node t is given in (2):

$$Entropy(t) = -\sum_i p(j|t)\log_2 p(j|t) \tag{2}$$

where $p(j|t)$ is related with frequency of category j at node t.

$$Gain = Entropy(t) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)\right) \tag{3}$$

Gain equation is depicted in (3) the parent node where t is divided into k partitions and n_i is number of data in partition i. A disadvantage is that bias of splitting; this can happen with many classes.

3.3 Gain Ratio (GR)

GR technique improves the problem of Info. The structure of method is created by using to-down design. GR was developed by Quinlan in 1986 and based on evaluation of information theory. Generally, probability, $(P(v_i))$, is to answer v_i, then the information or entropy (I) of the answer is given by [19]. *SplitINFO* is presented in (4) to resolve bias in Info. In (5), Info is adapted by using the entropy of the partitioning (*SplitINFO*). Thus, higher entropy partitioning is adjusted.

$$SplitINFO = -\left(\sum_{i=1}^{k} \frac{n_i}{n}\log\frac{n_i}{n}\right) \tag{4}$$

$$GR = \frac{\Delta\text{info}}{\text{SplitINFO}} \tag{5}$$

3.4 An Incremental Learning Algorithm based on Mahalanobis Distance (ILM)

The objective of ILM is incremental learning, covering supervised and unsupervised learning, hard and soft decisions [3] [20], Mahalanobis distance measurement, and Gaussian function [3]. ILM contains two steps: learning step and predicting step that is shown in Fig. 3. First, learning step consists of two learning algorithms (cover supervised and unsupervised learning) and an incremental learning algorithm. The learning algorithm is used to create a classification model and learning model using Mahalanobis distance and Gaussian membership function. On one hand, a new target class label can always learn by an incremental learning algorithm. Thus, the system prototype is updated regularly. However, the prototype is constructed using two parameters which are large degree of membership function and distance threshold of each target class label. The user defines both factors. In that case, the distance threshold is represented by d_{th}, where $0 < d_{th} < 1$. d_m is measurement distance among instance and If $d_m > d_{th}$, instance is dissimilarity with cluster. Conversely, instance has more similarity to cluster if $d_m \leq d_{th}$. Learning Model consists of system prototypes, which are W_P and W_T weight for supervised and unsupervised respectively.

Next, unknown data are inputted into the predicting step. This way, similarity or dissimilarity between unseen pattern and the system prototype is measured using Mahalanobis distance. In the case of incremental learning, it adds new prototypes if distance of unknown data are very high. The distance is calculated using Mahalanobis distance and Gaussian membership function. Soft decision characteristic is shown using Mahalanobis Gaussian RBF to estimate membership degree of each target class. Wining node will be the higher degree of membership function, this is called hard decision. Target class (W_P) of unknown data is determined using W_P label of winning node. Measuring distance between input (p) and target class is using the nonsingular covariance matrix (K) and calculating Mahalanobis distance (d) shown in (6) and (7). GR technique improves the problem of Info. The structure of method is created by using to-down design. GR was developed by Quinlan in 1986 and based on evaluation of information theory. Generally, probability, $(P(v_i))$, is to answer v_i, then the information or entropy (I) of the answer is given by [15]. *SplitINFO* is presented in (4) to resolve bias in Info.

$$K_{j,new} = K_{j,old} + aI_n \tag{6}$$

$$d = \sqrt{(p - W_p)^T K^{-1}(p - W_p)} \tag{7}$$

The membership degree of each cluster is calculated by Mahalanobis RBF which is represented in

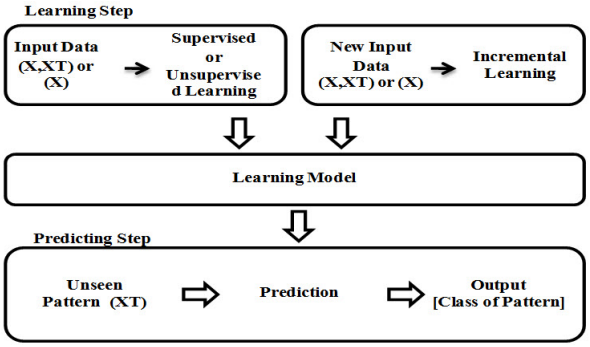

Fig.3: *Structure of the ILM Algorithm.*

(8). Maximum membership degree of winner cluster is found by fuzzy "OR" operator (e.g., max operator) in (9).

$$mem(p, W_p) = \exp\left(-\frac{(p - W_p)^T K^{-1}(P - W_p)}{2}\right) \tag{8}$$

$$\begin{aligned} winner &= \arg\max_i(men_i) \\ &= \arg\max_i(men_1 \vee men_2 \vee \ldots \vee men_n) \end{aligned} \tag{9}$$

4. EXPERIMENTS AND RESULTS

The combination between ILM and feature selection is used to confirm the validity for classifying gene expression data. Hence, this research focuses on only ILM supervised learning methods. Firstly, three public gene expression datasets are prepared, Colon Tumor (62×2001), Leukemia (72×7130), and Lymphoma (96×4026) shown in Table 1. Secondly, the discriminate gene of each dataset is selected based on filtering approaches which are Cfs, GR, and Info. Thirdly, the output from the previous step is transferred to ILM algorithm for classification. Moreover, this evaluation method is based on accuracy rate. Fig. 4 represents experiment design.
The details of operation are depicted below:

1. Data preprocessing: Replace the missing value and normalized data. Two public gene expression data sets contain missing value data which are Lymphoma and Leukemia dataset. Lymphoma dataset is higher missing value than Leukemia and shown in Fig. 5 and Fig. 6.

2. Filtering approach: Select the top-rank genes according to three feature selection methods (Cfs, GR, and Info) based on Colon Tumor, Leukemia, and Lymphoma Dataset that are shown in Table 2.

3. Training and testing datasets: Each dataset is separated into two groups, training (50%) and testing (50%) datasets. Training and testing datasets are used to construct learning and predicting, respectively.

4. Learning Model: transfer output of previous step (only training dataset) into ILM algorithm to create a learning model. In this case, covariance matrix and distance threshold is defined. Thus, the result of each filter approach is combined with ILM such as , CfsILM GRILM, and InfoILM.

5. Prediction: the performance evaluation of CfsILM, GRILM, and InfoILM based on accuracy rate of predicting target class and small subset of gene that has been discriminating power.

Table 1: *Detail of Gene Expression Datasets.*

Datasets	Instances	Attributes	Class Values
Colon Tumor	62	2001	Positive and Negative
Lymphoma	96	4026	GCL, ACL, DLBCL, GCB, NIL, ABB, RAT, TCL,FL, RBB, and CLL

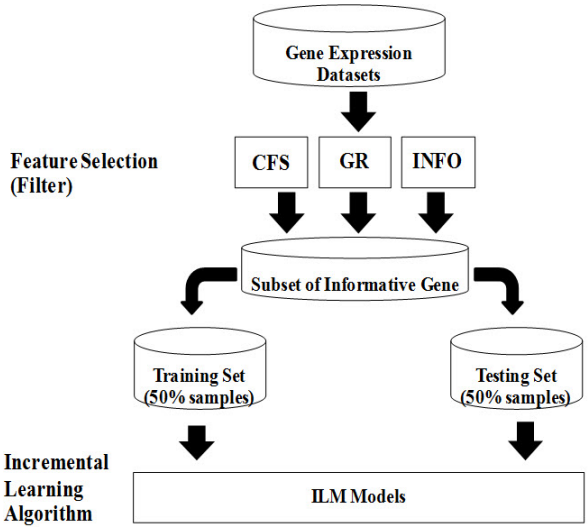

Fig.4: *Experiment Design for Classification Models Based on ILM Algorithm.*

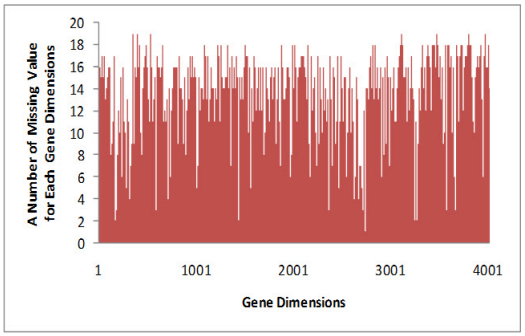

Fig.5: *Missing Value of Lymphoma Dataset.*

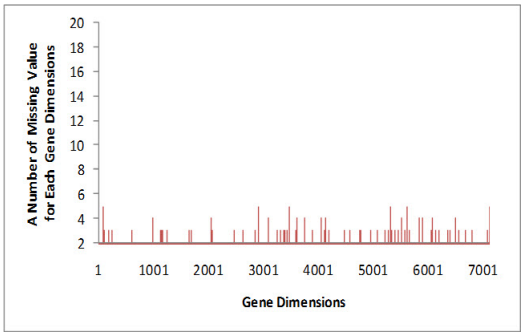

Fig.6: *Missing Value of Leukemia Dataset.*

Table 2: *The Result of Three FS Approaches..*

Datasets	Attributes	Feature Selection		
		CFs	GR	Info
Colon Tumor	2001	26	135	135
Leukemia	7130	75	874	874
Lymphoma	4026	269	1785	1785

4.1 The Result of Colon Tumor Dataset

The experimental results of feature selection based on Colon Tumor gene expression dataset are reported in Table 2. Using feature selections Cfs, GR, and Info, there were 26, 135, and 135 attributes respectively chosen as follows. Working together between feature selection and ILM algorithm represents important results CfsILM, InfoILM, and GRILM; it is higher accuracy than original ILM. As Fig. 7, these techniques can produce higher accuracy rate than traditional ILM technique.

4.2 The Result of Leukemia Dataset

Likewise, comparison of efficiency on original ILM and CfsILM, InfoILM, and GRILM based on Lymphoma dataset is depicted in Fig. 7 Meanwhile many features are chosen using feature selection Cfs, GR, and Info, which were 75, 874, and 874 represented in Table 2. The performance of integration CfsILM, InfoILM, and GRILM are represented in Fig.7.

4.3 The Result of Lymphoma Dataset

According to Lymphoma dataset, combining feature selection and ILM algorithm is higher efficiency than original ILM represented in Fig. 7. However, joining during two techniques; CfsILM, InfoILM and GRILM developed many attributes which were 269, 1785, 1785 respectively reported in Table 2.

Table 3: *The Result of Three FS Approaches..*

Datasets	Classification Models			
	ILM	InfoILM	GRILM	CfsILM
Colon Tumor	218.51	0.98	0.94	0.05
Leukemia	81,479.80	87.77	86.58	0.35
Lymphoma	2,000.00	1,876.09	1,989.46	9.30
Average	27,899.44	654.95	692.33	3.23

4.4 Comparison of Accuracy Rate and Time Efficiency

The experiment results shown all classification models increased accuracy rate over traditional ILM. This could be seen in Fig. 7. Obviously, that result shows CfsILM selected less number of attributes as Table 2 and get higher accuracy rate as Fig. 7 appropriate for classifying gene expression data. On the other hand, GRILM yielded lower efficiency when a subset of genes is selected. This exposed CfsILM to have high performance while GRILM had restricted capability. Nevertheless, Many attributes of the three gene expression datasets were eliminated and fed to the classification models. Furthermore, several dimensions on overall the dataset are chosen using GRILM and InfoILM model but a few attributes are selected using CfsILM model.

In case of time efficiency, three classification models (CfsILM, InfoILM, and GRILM) save time more than traditional ILM technique which is shown in Table 3. Especially, the average time of CfsILM is less than other classification techniques.

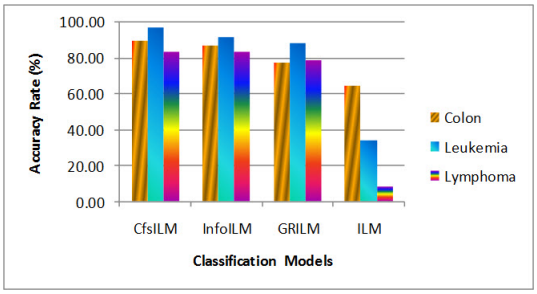

Fig.7: *The Result of Accuracy Rate Based on Classification Models.*

5. CONCLUSIONS

Genes can grow up into the tens of thousands, but most of them are not associated with discriminate power. Therefore, finding a subset of informative genes is very important for biological processing. This paper proposes classification models based on incremental learning algorithm and featuring selection on gene expression data. Three feature selection methods: Cfs, GR, and Info are combined with Incremental Learning Algorithm based-on Mahalanobis Distance (ILM) and three public gene expression datasets are used experimentally: Colon Tumor (62×2001), Leukemia (72×7130), and Lymphoma (96×4026). In the case of feature selection, classification models can eliminate many dimensions. For example, subsets of features based on Colon Tumor are decreased from 2001 into 26, 135, and 135, using CfsILM, GRILM, and InfoILM, respectively. Meanwhile, the accuracy rate of classification models is higher than the original ILM.

For instance, accuracy rates are calculated from experiments tested with three datasets that are expending from 64.52%, 34.29%, and 8.33% into 90%, 97.14%, and 83.33%, by CfsILM model, respectively. Therefore, the experiment result concludes CfsILM classification model joined with Cfs and ILM had not only higher performance over the other models but also average time saving 3.23 second.

6. ACKNOWLEDGEMENT

The major effort has acquired the whole-hearted support of Miss Maytiyanin Komkhao. Thank you very much for your source code and knowledge.

References

[1] Hongbo Xie, Uros Midic, Slobodan Vucetic, and Zoran Obradovic, *Hand Book of Applied Algorithms*, John Wiley and Sons, Inc., New York, 2008, ch. 5.

[2] Douglas H. Fisher, *Knowledge Acquisition Via Incremental Conceptual Clustering*, Machine Learning 2(2), Kluwer Academic Publishers, Boston, pp.139–172, 1987.

[3] M. Komkhao, S. Sodsee, and P. Meesad, "An Incremental Learning Algorithm Based-on Mahalanobis Distance for Unsupervised Learning," *Proceeding of 3th National Computer on Computing and Information Technology (NCCIT2007)*, pp.20-25, 2007.

[4] P. Lance, H. Ehtesham, and L. Huan, "Subspace Clustering for High Dimensional Data: A Review," *SIGKDD Explor. Newsl. 1931–0145*, vol. 6, pp.90–105, 2004.

[5] Pádraig Cunningham, *Dimension Reduction*, Technical Report UCD-CSI-2007-7., University College Dublin, pp.1-17, 2007.

[6] P. Saengsiri, Sageemas Na Wichian, Phayung Meesad, and Herwig Unger, "Comparison of Hybrid Feature Selection Models on Gene Expression Data," *Proceeding of 8th IEEE International Conference Knowledge Engineering*, pp.13–18, 2010.

[7] Mukherjee, S. and S. J. Roberts. "A Theoretical Analysis of Gene Selection," *Proceedings Computational Systems Bioinformatics Conference*, CSB, pp.131–141, 2004.

[8] Cheng-San, Y., C. Li-Yeh, Chao-Hsuan ke, and Cheng-Hong Yang, "A Hybrid Approach for Selecting Gene Subsets Using Gene Expression Data," *IEEE Conference on Soft Computing in Industrial Applications (SMCia'08)*, pp.159–164, 2008.

[9] Jaeger J., R. Sengupta , W. L. Ruzzo, "Improved Gene Selection for Classification of Microarrays," *8th Pacific Symposium on Biocomputing*, pp.53–64, 2003.

[10] Jin-Hyuk H. and C. Sung-Bae, "Cancer Classification Incremental Gene Selection Based on DNA Microarray Data," *IEEE Symposium Computational Intelligence in Bioinformatics and Computational Biology*, pp.70–74, 2008.

[11] Cheng-San, Y., C. Li-Yeh, Jung-Chike Li, and Cheng-Hong yang, "A Novel BPSO Approach for Gene Selection and Classification of Microarray Data," *IEEE International Joint Conference World Congress on Computational Intelligence Neural Networks (IJCNN2008)*, pp.2147–2152, 2008.

[12] Kerr, G., H. J. Ruskin, M. Crane, and P. Doolan, "Techniques for Clustering Gene Expression Data," *Computers in Biology and Medicine*, vol.38 no.3, pp.283–293, 2008.

[13] R. Ruiz, Jose C. Riquelme, and Jesus S. Aguilar-Ruiz, "Incremental Wrapper-based Gene Selection from Microarray Data for Cancer Classification," *Pattern Recognition*, vol. 39, pp.2383–2392, 2006.

[14] R. Dabney, "Classification of Microarrays to Nearest Centroids," *Bioinformatics*, vol. 21, no.22, pp.4148-4154, 2005.

[15] Q. Shen, W.-m. Shi, and W. Kong, "New Gene Selection Method for Multiclass Tumor Classification by Class Centroid," *Journal of Biomedical Informatics*, vol. 42, pp.59-65, 2009.

[16] M. Mohamad, S. Omatu, S. Deris, M. Misman, and M. Yoshioka, "A Multi-Objective Strategy in Genetic Algorithms for Gene Selection of Gene Expression Data," *Artificial Life and Robotics*, vol. 13, pp.410–413, 2009.

[17] Mark A. Hall, *Correlation-based Feature Selection for Machine Learning*, Doctor of Philosphy Thesis, Department of Computer Science, The University of Waikato Newzealand, pp.69–71, 1999.

[18] Pang-Ning Tan, Michael Steinbach, and Vipin

Kumar, *Introduction to Data Mining*, Addison Wesley, 2006, ch. 5.

[19] Quinlan, J. R, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp.81–106, 2006.

[20] G. G. Yen and P. Meesad, "An Effective Neuro-Fuzzy Paradigm for Machinery Condition Health Monitoring," *IEEE Transactions : Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, pp. 523–536, 2001.

Phayung Meesad Ph.D., assistant professor. He is an associate dean for academic affairs and research at the faculty of Information Technology, King Mongkut's University of Technology North Bangkok. He earned his MS and Ph.D. degrees in Electrical Engineering from Oklahoma State University, U.S.A. His researches are focused on data mining, hybrid intelligent system, optimization, and fuzzy logic algorithm.

Sageemas Na Wichian received her Ph. D. in Educational Research Methodology at Chulalongkorn University. She has been an educator for over ten years, taught in universities in the areas of diversity, Psychology and Research for Information Technology. She is currently working as a lecturer at King Mongkut's University of Technology North Bangkok, Thailand. Her researches are focused on advanced research methodology, Industrial and Organizational psychology.

Herwig Unger teaches Mathematics and Information Technology at the University of Hagen, Germany. He has 17 years of experience in many areas of the IT, e.g. simulation models, communication and networking, information management, grid computing, business intelligence and SCM. He focuses on communication and information distribution in networks for his research. Especially, the decentralize storage and data processing has been specially attended variously for his research projects.

Patharawut Saengsiri received Ph.D. in Information Technology at King Mongkut's University of Technology North Bangkok. He is currently working as an Information Technology Officer at Thailand Institute of Scientific and Technological Research (TISTR), Thailand. He had been scholarship student of Science and Technology Ministry of Thai Government. He is interested in A Development of an Incremental Hierarchical Clustering Algorithm for Gene Expression Data.