World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:6, 2008

Feature Subset Selection approach based on Maximizing Margin of Support Vector Classifier

Khin May Win, and Nan Sai Moon Kham

Abstract**—**Identification of cancer genes that might anticipate the clinical behaviors from different types of cancer disease is challenging due to the huge number of genes and small number of patients samples. The new method is being proposed based on supervised learning of classification like support vector machines (SVMs).A new solution is described by the introduction of the Maximized Margin (MM) in the subset criterion, which permits to get near the least generalization error rate. In class prediction problem, gene selection is essential to improve the accuracy and to identify genes for cancer disease. The performance of the new method was evaluated with real-world data experiment. It can give the better accuracy for classification.

Keywords**—**Microarray data, feature selection, recursive feature elimination, support vector machines.

I. Introduction

NOWADAYS, many researchers are investigating the class prediction methodology, especially for cancer classification. Recent advances in microarray technology allow scientists to measure the expression levels of thousands of features. New analytical methods are needed to be developed to identify the features of genes which have distinct signatures. Recently, Brown *et al.* applied a collection of supervised learning techniques to a set of microarray expression levels from yeast data [5]. They showed that an algorithm known as a support vector machine (SVM) [4] provides excellent classification performance. SVMs are members of a larger class of algorithms, known as kernel methods .Kernel methods are non-linearly mapped to a higher-order feature space by replacing the dot product operation in the input space with a kernel function $K(x, y)$. Mercer's theorem [8] shows that every positive semi-definite kernel function corresponds to the dot product operation in some higher-dimensional feature space. In this work, we construct an explicitly heterogeneous kernel function by computing separate kernels for each data type and summing the results. The resulting kernel incorporates prior knowledge about the heterogeneity of the data by accounting for higher-order correlations among features of one data type but ignoring higher-order correlations across data types. This heterogeneous kernel leads to improved performance with respect to an SVM trained directly on the concatenated data.

Feature subset selection refers the function that select the most relevant features to the classification task, removing the irrelevant ones. Two aspects are important in designing a feature subset selection: selection algorithm and selection criterion. Jain provided a useful taxonomy of selection algorithms [4] .But selection of genes shouldn't be independent of the algorithm to be used to construct the classifier. Evolutionary computation methods over ranking based feature selection have advantages because different combination of features is evaluated through generation of different individuals of feature population. In this paper, the proposed method is based on sequential selection approach which will be applied for classification task using support vector machines. The application of SVM is also found in a broad spectrum of technology in bioinformatics field [2].A new feature selection criterion function was proposed, based on the SVM's optimal margin. In the hope, reducing the non-informative feature provides a better presentation for the expected SVM's performance.

Manuscript received June 30, 2008. This work was supported by SDRC (Software Development and Research Center) in University of Computer Studies, Yangon. SDRC is a Research Center designated by Myanmar Science and Engineering Foundation and Ministry of Science & Technology.

Khin May Win is now doing research for Ph.D (IT) about Microarray Data for Cancer Classification Project at University of Computer Studies, Yangon, Myanmar (e-mail: winn.km05@gmail.com).

II. Related Work

A. Feature Selection based SVMs

It is essential to efficiently analyze the microarray data because the amount of DNA microarray data is usually very large. The analysis of DNA microarray data is divided into branches such as clustering, classification, gene identification, and gene regulatory network modeling [1]. Many machine learning and data mining methods have been applied to solve them .Specifically, gene selection is used to identify genes most relevant to sample classification, such as normal and cancerous gene samples. The most common gene selection approach is so-called gene ranking. Among various gene selection methods, support vector machine-based recursive feature elimination (SVM-RFE) has become one leading method and is being widely used. The support vector machine learning algorithm is such a technique. Moreover, since the gene selection is based on an SVM-classifier, a subset of genes that yields high classification performance can be identified [6]. Eliminate the irrelevant features (genes) using the weight vector of the hyperplane constructed by the samples on the margin i.e. support vectors, based on the

maximum margin criterion.

B. Support Vector Machines

Support Vector Machines (SVM) have arisen many attentions in pattern recognition field. SVMs scale well and have been used successfully with large training sets in the domains of text categorization and image recognition [7]. Furthermore, in this paper, we demonstrate that SVMs can learn from heterogeneous data sets for separable in feature space. With an appropriate kernel function, the SVM learns from a combination of two different types of feature vectors. In most cases, the resulting trained SVM provides as good or better gene functional classification performance than an SVM trained on either data set alone. SVM builds up a hyperplane as a decision surface in such a way to maximize the margin of separation between positive and negative examples. Depending on the Vapnik-Chervonenkis (VC) dimension. Constructing an optimal hyperplane is equivalent to finding all nonzero support vectors and a bias. Several machine learning techniques have been previously used in classifying gene expression data, including Fisher linear discriminated analysis [4], *k* nearest neighbor [9], decision tree, multi-layer perceptron [5,15],support vector machine[4,5]. Also, many machine learning techniques were have been used in clustering gene expression data [11]. They include hierarchical clustering [2], self-organizing map [8], and graph theoretic approaches are also appeared by Hartuv and Sharan. On the other hand, it would be difficult to generate compact size feature subsets. These are motivated us to design a method that can reduce the number of features dimension in different assuming.

III. Feature SubSet Selection based on Optimal Margin

Let Y be the set that comprises all the n origin features, and $X(X \subseteq Y)$ a selected subset of Y, with d features. *J(X)* is a function that determines how good the subset X is, by a certain criterion.

$$J(X) = \max_{Z \subseteq Y, |Z|=d} J(Z) \qquad (1)$$

J is the various criterion functions to measure the quality of the subset of features. Based on the selection criterion, the methods of feature subset selection can be categorized into two approaches: Filter and Wrapper. The filter methods employ the intrinsic properties of data , such as class separability [5].The wrapper methods [6]evaluates the quality of the subset based on the classification rate of a classifier. Generally, the wrapper method achieves better performance than that of filter method, but the computational cost is expensive due to the absence of the test data.

A. Ordinary Margin of SVMs Hyperplane

Support vector machine (SVM) estimates the function classifying the data into two classes by Vapnik Theory [15]. SVM builds up a hyperplane as the decision surface in such a way to maximize the margin of separation between positive and negative examples. For the separable data points, how much closer to the hyperplane that means the average absolute distance of all samples to the hyperplane.

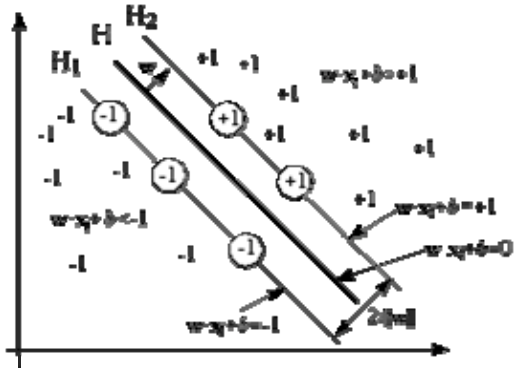

Fig. 1 Optimal hyperplane with maximized margin

Ordinary margin measures the distances between the instances in the space and the decision boundary induced by the SVMs. In this approach, we maximize the margin to obtain the optimal hyperplane in the feature space. In order to distinguish the ordinary margin of hyperplanes from the proposed criterion.

Let $x_i \in R^n (i = 1,...,l)$ be a feature vector, where l is the number of samples and $y_i \in \{-1,1\}$ the class assigned to each sample. The discrimination function of SVM is given by:

$$f(x) = w.\Phi(x) + b \qquad (2)$$

The weight vector w is obtained as follows:

$$w = \sum_{i=1}^{l} \alpha_i y_i \Phi(x_i) \qquad (3)$$

where α are the Lagrange multipliers.

Using the weight vector, w, the geometrical interpretation of the ordinary margin is defined as:

$$NM = \frac{1}{\| w \|} \qquad (4)$$

and ||w|| is given by:

$$\| w \|^2 = \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \Phi(x_j) \qquad (5)$$

As $\Phi(x_1)\Phi(x_2) = K(x_1, x_2)$, by substituting equation (5) in equation (4), it is possible to rewrite the ordinary margin as follows:

$$NM = \left(\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_1, x_2) \right)^{-\frac{1}{2}} \qquad (6)$$

In this study, the Gaussian Kernel was used, which is defined as:

$$K(x_1, x_2) = \exp\left(-\frac{\| x_1 - x_2 \|^2}{2\sigma^2}\right) \quad (7)$$

In this work, the weight vector norm ||w|| is used as criterion for subset evaluation as follows:

$$\left| \| w \|^2 - \| w^{(i)} \|^2 \right| =$$

$$\frac{1}{2}\left| \sum_{j,k=1}^{l} \alpha_i \alpha_j y_i y_j K(x_j, x_k) - \sum_{j,k=1}^{l} \alpha_i^{(i)} \alpha_j^{(i)} y_i y_j K^{(i)}(x_j, x_k) \right| \quad (8)$$

where $K^{(i)}$ and $\alpha_j^{(i)}$ are defined respectively. As the Kernel function values and the Lagrange multipliers obtained in the absence of the i[th] feature.

When the worst feature Kernel is found. It is removed, which gives a score, named SR, for that subset, calculated from $K^{(k)}$ and $\alpha_j^{(k)}$ .This method is based on the feature elimination.

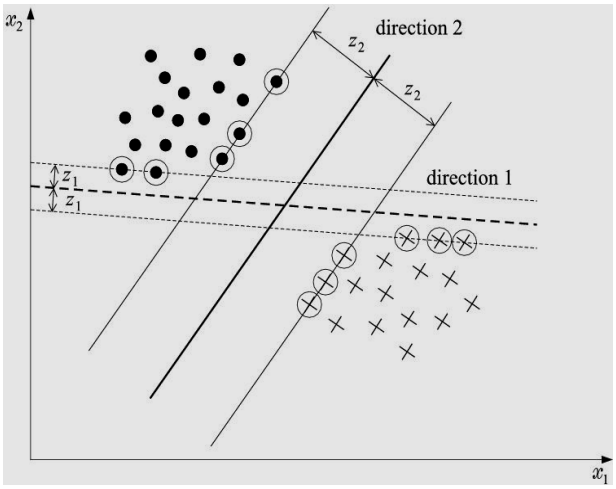

Fig. 2 Maximizing Margin width on both directions

The use of Maximized Margin (MM) as criterion for subset evaluation and its behaviors in the presence of misclassified data. Even the features are ranked; the maximal value of SR does not correspond to the best feature subset, still requiring a monitoring of the classifier accuracy.

B. Sequential Backward Selection for Maximizing Margin

By applying the idea of the maximizing ordinary margin. Furthermore, similar confidence subsets will be discriminated by the value of the ordinary margin itself. Hence, it is expected that this new measurement for maximum margin gives better results in the feature subset selection task. The maximizing margins apply the sequential backward selection method.

For the sake of simplicity ,the proposed algorithm will be referred to as Sequential Backward selection using Maximized margin (SBS-MM).The selection strategy used in this algorithm is developed based on the approach of Marill and Green, which work in top-down fashion [4].The selection process starts from a full set of feature , then removes sequentially the most irrelevant ones. To find the most irrelevant feature of the current surviving subset, one of the features (e. g the ith feature) is removed and the size of margin is calculated. This is denoted as MM (i) ,i .e. the maximized margin without ith feature. The ith feature is returned to the subset, this procedures work until the features are over. Finally, the most irrelevant feature, which its removal produced the greatest value of maximized margin, can be found. The procedure is repeated until all of the features are removed. By monitoring the identification of the best subset which has the maximum generalization performance of the generated ranking of the each feature.

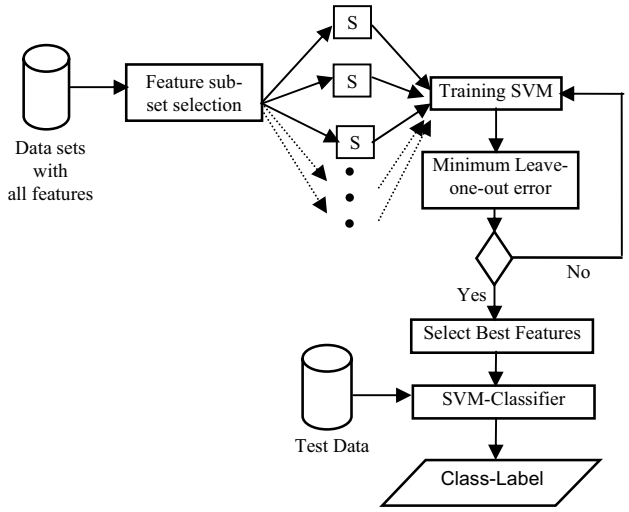

Fig. 3 System Flow with Sequential Backward Feature Selection

Algorithm

Step 1: Initialize:

Subset of surviving features *s =[1,2,...,n]*

Step 2: Repeat

For $\forall s_i \in s(1 \le i \le |\, s\,|)$

Do train the SVM classifier without i[th] feature

Do compute $Ji=MM^{(i)}$

/*$MM^{(i)}$ is the maximized margin without ith feature*/

Find the worst feature k

$$k = m \,|\, J_m = \arg\max_q (J_q)$$

Remove the feature that maximizes *MM*

s *= [1,… ….,k-1,k+1,…..n]*

Step 3: Until ***s*** is empty.

Finally, the most irrelevant features, which its removal produced the greatest value of ordinary margin can be found. It is expected to be possible to identify the best subset which has the best generalization performance of feature ranking. Since the randomly-generated attributes can be correlated with the response, they cannot always be eliminated. Thresholding based on the maximum weight rather than the average weight on the random variables.

IV. Experimental Result

A gene subset is evaluated by its accuracy on the training data and the number of genes selected in it. In this method, we can calculate the evaluation measure with function

$$ER(x) = w * A(x) + (1-w) * (1 - NGS(x)/n) \quad (9)$$

where A(x) is the accuracy on training data using the selected genes in x, NGS(x) is the number of genes selected in x and $w \in [0,1]$ is the assigned value for accuracy. In our experiments, we give more emphasis on accuracy rather than on number of selected genes.

A. Accuracy Estimation by Support Vector Machines

An SVM is a maximum-margin hyperplane that lies in some space. Given training examples labeled as either"+1" or"-1", training examples such that the distance from the closest examples (the margin) to the hyperplane is maximized. In Leave–One-Out-Cross-Validation (LOOCV), one sample from the training set is excluded, and rest of the training samples are used to build the classifier. Then the classifier is used to predict the class of the left out one, and this is repeated for each sample in the training set. The LOOCV estimate of accuracy is the overall number of correct classifications, and then the classes of the samples are predicted one by one by taking the gene expressions of the selected genes.

B. Data Sets

For our experiments, we have chosen three microarray data sets of cancer research. These include Lung Carcinoma, Brain Cancer and Prostate Cancer. In table 2, #Genes denotes the number of genes that are left after preprocessing **.**

(1)Leukemia dataset

It contains 38 samples from 2 classes of Leukemia: 27 acute lymphoblastic (ALL) and 11 acute myeloid (AML).Other 34 samples consisting of 20ALL and 14AML are used as an independent test set.

(2)Prostate Cancer dataset

It contains 102 samples from 2 classes: 50 normal tissue samples and 52 prostate tumor samples.

(3)Breast Cancer dataset

It contains 76 samples from 2 classes of survival.33 poor prognosis and 43 good prognosis. Other 19 samples with 12 poor prognosis and 7 good prognosis are used as independent test set.

C. Preprocessing

For the Prostate cancer and Leukemia, each sample was standardized to zero mean and unit variance across genes. For the Breast cancer dataset, each sample was standardized after filtering of genes. As a baseline gene selection criterion, top ranked genes with the largest ratios were used for classification. In this case, we apply the nearest mean classifier to find the effective cancer classification.

TABLE I
Microarray Data Set Used in Experiments

Data Set	#Genes	Classes	#Samples
Leukemia	7129	AML&ALL	34
Breast Cancer	4918	Poor &Good prognosis	76
Prostate Cancer	2600	Normal & Tumor	102

TABLE II
Best Results Obtained by our Gene Selection Method with α =1.4 and C = {0.01, 0.1, 1,10, 100}

Data Set	Training Accuracy	Minimum number of selected genes
Leukemia	90.2% (#Genes=98)	48
Breast Cancer	90.48% (#Genes=66)	15
Prostate Cancer	96% (#Genes=47)	24

TABLE III
Performance Comparison for Binary Data Sets

Data Set	Selected genes	SVM-RFE (H) (%)	SVM-RFE (S) (%)
Leukemia	10 to 48	5.6 ± 1.6	6.5 ± 1.6
Breast Cancer	10 to 15	45.0 ± 1.4	35.6 ± 0.6
Prostate Cancer	10 to 24	10.5 ± 1.5	9.8 ± 0.8

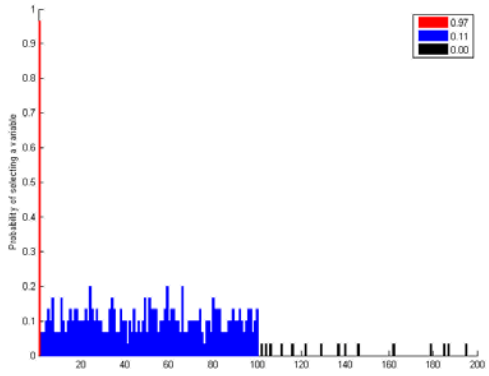

Fig. 4 Feature selection when C is small (≤ 0.01)

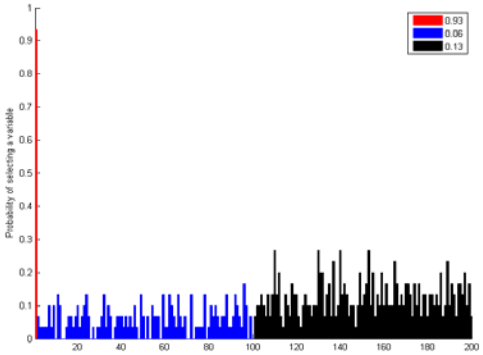

Fig. 5 Feature selection when C is large (≥ 0.1)

TABLE IV
RECOGNITION RATE WITH FEATURE SELECTION METHOD

Data	SVM		KNN	
	Linear	RFE	Cosine	Pearson
Leukemia	79.4	79.4	79.1	74.1
Breast Cancer	77.4	71.9	73.5	73.2
Prostrate Cancer	79.89	79.95	76.0	76.9

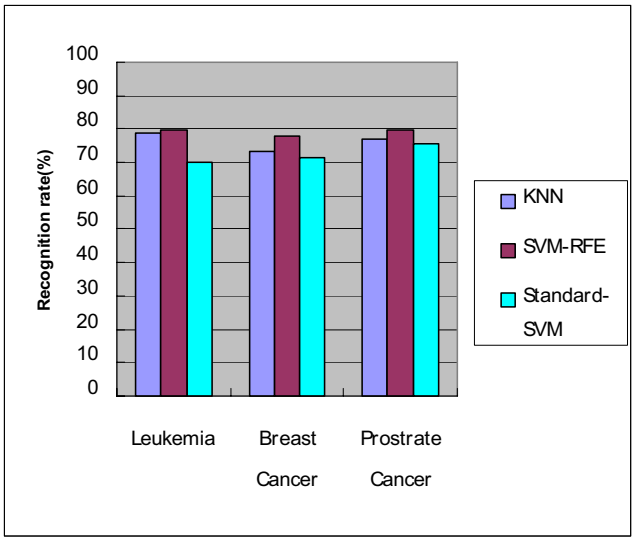

Fig. 6 Average Performance of the best classification with feature selection

The average error rates show at Table III. After feature selection, the number of genes are between 100 and 10 with respect to the Confident Margin parameter(C).The overall result on each data set with the setting of the parameters as described. Our work has done on training data set and calculates accuracy through leave one out cross validation. The average overall accuracy and the number of genes selected are shown in Table II. Interestingly, the suitable gene subset by maximizing margin that produces better accuracy and including top ranked gene. From this, we can infer that there may exit some kind of correlations among the selected genes. On the other hand, when we take a single gene from each subset, the correlation breaks down and it does not produce better accuracy on the data set.

The results of recognition rate on the test data are as shown in Table IV. Column is the list of comparing feature selection methods with KNN. Similarity measures used in KNN are Pearson correlation coefficient and Cosine coefficient. The difference of performance in data sets might be caused by the characteristics of data.

V. CONCLUSION

In this study, a new feature subset selection algorithm for classification task using SVMs was developed. The proposed method was assumed that very few features are needed to classify, the given samples and smallest subset may provide more insight into the data. During fitness calculation of a gene subset, minimum number of genes and maximization of classification accuracy have been scalarized. The margin of the SVM is base to the evaluation criterion of selected feature subset that measurement is the new selection criterion. In terms of dimensionality reduction, the best accuracy we get by starting from the original set and reduces the irrelevant features in each individual.

ACKNOWLEDGMENT

This research was supported by SDRC (Software Development and Research Center) in University of Computer Studies, Yangon. SDRC is a Research Center designated by Myanmar Science and Engineering Foundation and Ministry of Science & Technology.

REFERENCES

[1] A.Jain and D.Zongker, "Feature Selection: Evaluation, application and small performance", IEEE Transaction on Pattern Analysis and Machine Intelligence,vol.19,no.2pp.153-158,1997.

[2] C.Emmanouilidis, A.Hunter,and J.MacIntyre,"A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator", in Proceedings of the 2000 Congress on Evolutionary Computation(CEC00).

[3] Cancer Program Data Set [http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi]

[4] Eisen, M.B and brown, P.O.(1999); DNA arrays for analysis of gene expression. Methods Enzymbol, 303: 179-205

[5] Furey TS, Cristianini N, Duffy N, Bednarski DW,Schummer M, Haussler D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**:906-914.

[6] Kim,H.D.and Cho,S.-B(2000):Genetic optimization of structure-adaptive self-oranization map for efficient classification. *Proc. of International Conference on Soft Computing,*34-39,World-Scientific Publishing.

[7] K.M. Win and Kham N.S.M, "Minimizing Essential Set Based Feature selection for Cancer Classification", ICCA2008, Yangon, Myanmar, Feb 14-15, 2008

[8] M.PBrown,W.N.Grundy,D.Lin,N.Cristianini, C.W. Sugnet,J. Ares,Manuel, and D.Haussel"Support Vector machine classification of microarray gene expression data",Universith of California,Santa Cruz,Tech.June 1999.

[9] P.Larranga and J.Lozano. Estimation of distribution Algorithm:A new Tool for Evolutionary Optimization. Kluwer Academic Publishers, Boston, USA, 2001

[10] R.Kohavi and G.H.John,"Wrappers for feature subset selection" Artificial Intelligent,vol.97,no.1-2,pp.273-324,1997.

[11] R.Gilad-Bachrah,A.Navot,N.Tishby,"Margin based feature selection theory and algorithms"in proceeding of the 21st International Conference on Machine Learning(ICML04).New York :ACM Press,2004.

[12] Shamir, R. and Shanran,R (2001):Algorithmic approaches to clustering gene expression data.*Current Topic in Computational Biology.In Jiang,T.,Smith,T.,Xu,Y.and Zhang .M.Q.(eds),MIT press*

[13] T.Marill and D.Green,"On the effectiveness of reporters in recognition systems",IEEE Transactions on Information Theory,vol.9,pp.11-17,1999

[14] T.Paul and H.Iba.Selection of the most useful subset of genes for gene expression –based classification. In Proceeding of the 2004 Congress onEvolutionary Computation (CEC 2004),pages 2076-2083,Portland,Oregon,USA,2004.

[15] V.N Vapnik. The Nature of Statistical Learning Theory .Springer, New York, 1995.

[16] Xu J, Zhang X, Li Y: Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR**.** In *Proceedings of the International Joint Conference on Neural Networks: 15–19 July 2001* Washington, DC, IEEE;2001:1486-1491.

[17] Zhu J, Hastie T: Classification of gene microarrays by penalized logistic regression**.** *Biostatistics* 2004, **5:**427-443.

Khin May Win was graduated from University of Computer Studies, Mandalay at 2000 in Myanmar. She received her M.I.Sc Degree from University of Computer Studies at Mandalay in Myanmar, in 2002.After graduation she worked as a tutor in University of Computer Studies (Taunggyi) for three years .Khin May Win is a member of Myanmar Computer Foundation (MCF) since 2001.The major field of study for Khin May Win is numerical and statistical studies in Machine Learning of Bioinformatics field.

She worked as a tutor between 2003 and 2004 at University of Computer Studies at Taunggyi in Myanmar. She worked as an Assistant Lecturer between 2005 and 2008 at University of Computer Studies, Yangon at Myanmar. She has been working at University of Computer Studies, Yangon since 2005. .At present, she has two Conference Papers.One is for the BGRS '2008, Novosibirisk, Russia and the other one for ICCA 2008, Myanmar.