Cluster Based Hybrid Niche Mimetic and Genetic Algorithm for Text Document Categorization

A. K. Santra[1], C. Josephine Christy[2], B. Nagarajan[3]

[1] **Dean, CARE School of Computer Applications , Trichy – 620 009, India**.
[2] **Research Scholar, Bharathiar University, Coimbatore – 638401, India.**
[3] **Associate Professor, Bannari Amman Institute of Technology, Sathyamangalam – 638401, India.**

Summary

An efficient cluster based hybrid niche mimetic and genetic algorithm for text document categorization to improve the retrieval rate of relevant document fetching is addressed. The proposal minimizes the processing of structuring the document with better feature selection using hybrid algorithm. In addition restructuring of feature words to associated documents gets reduced, in turn increases document clustering rate. The performance of the proposed work is measured in terms of cluster objects accuracy, term weight, term frequency and inverse document frequency. Experimental results demonstrate that it achieves very good performance on both feature selection and text document categorization, compared to other classifier methods.

Keywords:
Document categorization, Feature Selection, Niche Mimetic Algorithm, Genetic Algorithm

1. Introduction

In the last ten years, content-based document management tasks have gained a prominent status in the information system field, due to the increased availability of documents in digital form and the ensuring need to access them in flexible ways [2]. Clustering is an important task in unsupervised learning. The essence of the clustering problem is to partition a set of objects into an a priori unknown number of clusters while minimizing the within cluster variability. Then it is maximizing the between cluster variability. Data clustering is a common technique for statistical data analysis and has been used in a variety of engineering and scientific disciplines such as biology.

While a wide range of classifiers have been used, virtually all of them were based on the same text representation, bag of words, where a document is represented as a set of words appearing in this document. Features used to describe a word are usually the ones which express whether the word appears in a document or how frequently this word appears. Above all, while the frequency of a word expresses the intuition that the more frequent, the more important, the compactness of the appearances of a word shows that the less compact, the more important and the position of the first appearance of a word shows that the earlier, the more important. Experiments suggest that the distributional features are useful for text categorization.

Document clustering groups similar documents into clusters on the basis of their contents. The documents in the resultant clusters exhibit maximal similarity to those in the same cluster and, at the same time, share minimal similarity with documents from other clusters. In addition, existing monolingual document clustering techniques can be classified broadly into non-LSI-based and LSI-based approaches. Memetic algorithms (MA) represent one of the recent growing areas of research in evolutionary computation. The term MA is now widely used as a synergy of evolutionary or any population-based approach with separate individual learning or local improvement procedures for problem search.

Here we suggest a unified criterion for simultaneous clustering and feature selection based on a well-known scatter separability index. A GA-based evolutionary procedure is then proposed to optimize the criterion. In order to allow simultaneous clustering and feature selection without the number of clusters being known a priori, a composite representation is devised to encode both feature election and cluster centers with a variable number of clusters. As a consequence, the crossover and mutation operators are suitably modified to tackle the concept of composite chromosomes with variable lengths. Additionally, we hybridize the proposed procedure with local search operations, which are introduced to refine the feature selection and cluster centers, respectively. These local searches move solutions toward local optima and allow a significant improvement in the computational efficiency. Finally, a niche method is integrated with the resulting hybrid GA to preserve the population diversity and prevent premature convergence.

2. Literature Review

S. Areibi and Z. Yang [1] have proposed several local search operations to effectively design an MA for simultaneous clustering and feature selection. which incorporate local searches with traditional GAs, have been proposed and applied successfully to solve a wide variety of optimization problems. These studies show that pure GAs are not well suited to fine tuning structures in complex search spaces and that hybridization with other techniques can greatly improve their efficiency. J. Shi and J. Malik [2] and S. Wu *et al.*[3] have proposed about data clustering is a common technique for statistical data analysis and has been used in a variety of engineering and scientific disciplines such as biology (genome data). Y. Zhao and G. Karypis [4] have proposed the purity of a cluster represents the fraction of the cluster corresponding to the largest class of documents assigned to that cluster; thus, the purity of the cluster.

One way of approaching this challenge is to use stochastic optimization schemes, prominent among which is an approach based on genetic algorithms (GAs). The GA is biologically inspired and embodies many mechanisms mimicking natural evolution. It has a great deal of potential in scientific and engineering optimization or search problems. Recently, hybrid methods [2], [5], [6], which incorporate local searches with traditional GAs, have been proposed and applied successfully to solve a wide variety of optimization problems. These studies show that pure GAs are not well suited to finetuning structures in complex search spaces and that hybridization with other techniques can greatly improve their efficiency. GAs that have been hybridized with local searches are also known as memetic algorithms (MAs) [7].

Traditional GAs and MAs are generally suitable for locating the optimal solution of an optimization problem with a small number of local optima. Complex problems such as clustering, however, often involve a significant number of locally optimal solutions. In such cases, traditional GAs and MAs cannot maintain controlled competitions among the individual solutions and can cause the population to converge prematurely [8].

To improve the situation, various methods [9], [10] (usually called niche methods) have been proposed. The research reported shows that one of the key elements in finding the optimal solution to a difficult problem with a GA approach is to preserve the population diversity during the search, since this permits the GA to investigate many peaks in parallel and helps in preventing it from being trapped in local optima.

GAs are naturally applicable to problems with exponential search spaces and have consequently been significant source of interest for clustering [6]. For example, in [11] proposed the use of traditional GAs for partitioned clustering. These methods can be very expensive and susceptible to becoming trapped in locally optimal solutions for clustering large data sets. Tsai et al. [6] introduced hybrid GAs by incorporating clustering-specified local searches into traditional GAs.

In contrast to the methods proposed in [3] and [5], clustering based on hybrid GAs can be more efficient, but these techniques can still, however, suffer from premature convergence. Furthermore, all of the above methods may exhibit limited performance, since they perform clustering on all features without selection. GAs have also been proposed for feature selection [3], [6]. However, they are usually developed in the supervised learning context, where class labels of the data are available, and the main purpose is to reduce the number of features used in classification while maintaining acceptable classification accuracies.

The second (and related) theme is feature selection for clustering, and feature selection research has a long history, as reported in the literature. Feature selection in the context of supervised learning [7], [6], adopts methods that are usually divided into two classes [2], [3] filters and wrappers based on whether or not feature selection is implemented independently of the learning algorithm. To maintain the filter/wrapper distinction used in supervised feature selection, we also classify feature selection methods for clustering into these two categories based on whether or not the process is carried out independently of the clustering algorithm.

The filters in clustering basically preselect the features and then apply a clustering algorithm to the selected feature subset. The principle is that any feature carrying little or no additional information beyond that subsumed by the remaining features is redundant and should be eliminated.

3. Document Clustering

Document clustering is very much for categorizing documents into meaningful groups. The usefulness of categorization is fully appreciated with labeling the clusters with the relevant feature words or phrases which describe various text document associated with them. A highly accurate key phrase extraction algorithm, called Core Phrase is proposed for this particular purpose.

Core Phrase works by building a complete list of phrases shared by at least two documents in a cluster. Phrases are assigned scores according to a set of features calculated from the matching process. The candidate phrases are then ranked in descending order and the top L phrases are output as a label for the cluster. While this algorithm on its own is useful for labeling document clusters, it is used to produce cluster summaries for the collaborative clustering algorithm.

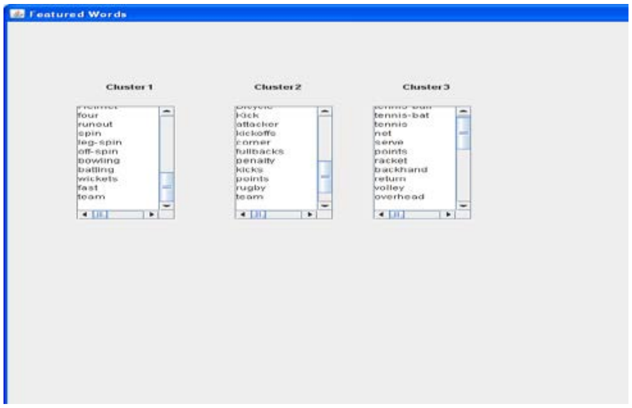
Cluster key phrase summaries are exactly what is used to succinctly inform remote nodes of the content of local document clusters, which in turn is used to judge the similarity between remote data and local clusters. A distributed version of this algorithm is also used in the hierarchically-distributed clustering algorithm (described below) to produce summaries for the globally distributed clusters.

- Key phrase extraction can be applied to a single document for labeling the document; this is mainly used in generating metadata (e.g. title, description, keywords) that can be associated with the document.
- A centralized document cluster can be summarized and labeled using key phrase extraction.
- Distributed document clusters in a flat peer-to-peer network can be summarized. Cluster summaries can be exchanged between peers to facilitate collaborative clustering.
- Distributed document clusters in a hierarchical peer-to-peer network can be summarized. Cluster summaries can be accessed at different levels of the hierarchy, thus providing variable scope of summaries ranging from specific to broad.

Evaluation of the accuracy of Core Phrase shows that it can accurately extract those phrases that match the manually labeled topic of clusters, and is able to rank those matching phrases in the top two or three key phrases.

Document clustering is used to organize a large document collection into distinct groups of similar documents. It discerns general themes hidden within the corpus. Applications of document clustering go beyond organizing document collections into knowledge maps. This can facilitate subsequent knowledge retrievals and accesses. Document clustering, shown in Fig 1. for example, has been applied to improve the efficiency of text categorization and discover event episodes in temporally ordered documents. In addition, instead of presenting search results as one long list, some prior studies and emerging search engines employ a document clustering approach to automatically organize search results into meaningful categories and thereby support cluster-based browsing.

Fig 1. Document Clustering



Cluster 1	Cluster 2	Cluster 3
four	Kick	tennis-bat
runout	attacker	tennis
spin	kickoffs	net
leg-spin	corner	serve
off-spin	fullbacks	points
bowling	penalty	racket
batting	kicks	backhand
wickets	points	return
fast	rugby	volley
team	team	overhead

Various document clustering techniques have been proposed, but most deal with monolingual documents (i.e., all target documents are written in the same language). However, the globalization of business environments and advances in Internet technology often cause an organization to maintain documents in different languages in its knowledge repositories. Evidently, organizations face the challenge of multilingual document clustering (MLDC). Such MLDC requirement is also prominent in other scenarios. For example, with advances in cross-lingual information retrieval (CLIR) technology, many search engines now offer a functionality that retrieves, for a user query expressed in one language, relevant documents in different languages.

In this case, to facilitate cluster based searching, it would be preferable if the search engine were capable of clustering search results in different languages into distinct categories, each of which contains documents similar in their contents.

In our work feature selection is carried out to categories document with clustering method. The feature selection process is accomplished with integrated niche memetic and genetic algorithm which are explained in the following sub sections.

3.1 Feature Selection

Feature selection is important for clustering efficiency and effectiveness because it not only condenses the size of the extracted feature set but also reduces any potential biases embedded in the original (i.e., non-trimmed) feature set . Previous research commonly has employed feature selection metrics such as TF (term frequency), TF×IDF (term frequency × inverse document frequency), and their hybrids. A sample example is shown in Table 1.

Unlike the non-LSI-based document clustering approach, which typically involves a feature selection phase, the LSI-based approach to clustering monolingual documents employs LSI to reduce the dimensions and thereby improve both clustering effectiveness and efficiency. Its process generally commences with feature extraction, followed by document representation.

Name	Frequency	Inverse Frequency	Weight
tennis ball	0.33	3.0	1.0
tennis-bat	0.33	3.0	1.0
Tennis	0.33	3.0	2.0
Net	0.33	3.0	1.0
Serve	0.33	3.0	1.0
Points	0.33	3.0	1.0
Rocket	0.33	3.0	1.0
backhand	0.33	3.0	1.0
Return	0.33	3.0	1.0
Volley	0.33	3.0	1.0
Overhead	0.33	3.0	1.0

Table 1: Term Frequency - Tennis

4.1. Niching Memetic Algorithm

In our hybrid scheme of text document categorization, we used a niche MA for simultaneous clustering and feature selection by optimizing the unified criterion. This algorithm works with variable composite chromosomes, which are used to represent solutions. The operation of the algorithm consists of using a niche selection method for selecting pairing parents for reproduction, performing different genetic operators on different parts of the paired parents, applying local search operations (i.e., feature add and remove procedures and one step of K Means) to each offspring, and carrying out a niche competition replacement.

The evolution is terminated when the fitness value of the best solution in the population has not changed for g generations. The output of the algorithm is the best solution encountered during the evolution.

The flow of the algorithm is given as follows:

Step 1: initialize p sets of solutions randomly which encode both feature selection and cluster centers

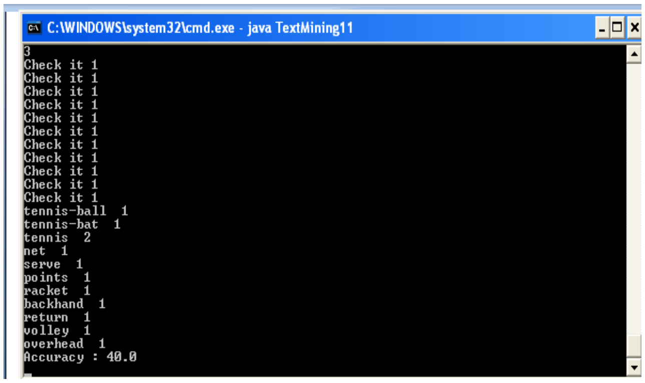
Step 2: Calculate unified criterion J2 and set its fitness value as f ¼ J2.

Step 3: Repeat the following steps until the stopping criterion is met:

i. Select pairing parents until p=2 parent pairs are selected.
ii. Generate intermediate offspring by applying different genetic operators on the different parts of the paired parents.
iii. Apply feature add and remove procedures to the offspring.
iv. Run one step of K Means on the offspring.
v. Pair the offspring with the most similar solution found during a restricted competition replacement.
vi. Calculate J2 according to (4) for each of the offspring. If the fitness of the offspring is better than its paired solution, then the latter is replaced.

Step 4: Provide the feature subset and cluster centers of the solution

The accuracy rate can be calculated by utilizing the Nichie Memetic algorithm is shown in the Fig 2.



```
3
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
tennis-ball  1
tennis-bat  1
tennis  2
net  1
serve  1
points  1
racket  1
backhand  1
return  1
volley  1
overhead  1
Accuracy : 40.0
```

Fig: 2. Accuracy Rate of Nichie Memetic Algorithm

4.2. GA Algorithm

In a genetic algorithm, a population of strings (called chromosomes or the genotype of the genome), which encode candidate solutions (called individuals, creatures, or phenotypes) to an optimization feature selection of text from documents, evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations.

For the experiments, the system set genetic parameters as follows:

Generation limit = 100

Population size = 100

Cross-over fraction = 0.8

Mutation fraction = 0.1

Reproduction fraction = 1

Top N selection = 100

In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached. The flow of the algorithm is given as follows:

step1: Initialize the Population

step2: Evaluate fitness of each member

step3: Reproduce with fittest members

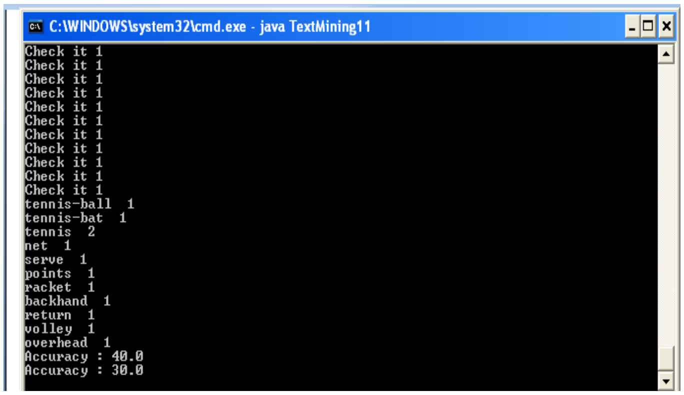
step4: Introduce random mutations in new generation

step5: Continue (2)-(3)-(4) until pre-specified number of generations are complete

Finally the hybrid niche memtic and genetic algorithm joins the process of feedback on the traditional foundation frame "training → Categorizing " algorithm. It expands the algorithm process as "Training → Categorizing feedback judgment → feedback". This kind of method is more close the real meaning machine learning. It show that the proposed hybrid algorithm has certain degree cognition self- determination in text document categorization using clustering methods.

The accuracy rate can be calculated by utilizing the Genetic Algorithm is shown in the Fig 3.

Fig 3: Accuracy Rate of GA Algorithm



```
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
Check it 1
tennis-ball  1
tennis-bat  1
tennis  2
net  1
serve  1
points  1
racket  1
backhand  1
return  1
volley  1
overhead  1
Accuracy : 40.0
Accuracy : 30.0
```

5. Experimental Result and Discussion

In the experiment, we used Reuters-578, which has 178 documents collected from the Reuters newswire, as training sample set. Of the 35 categories in Reuters 578, only the most populous 10 are used. In data pre-processing, we applied stop word removal and tfc feature selection, and removed the commoner morphological and inflexion endings from words using The Porter Stemming Algorithm. Each category is employed as the positive class, and the rest as the negative class. For each dataset, 30% of the documents are randomly selected as test documents, and the rest are used to create training sets as follows: γ percent of the documents from the positive class is first selected as the positive set P. The rest of the positive documents and negative documents are used as unlabeled set U. We range γ percent from 10% - 50% to create a wide range of scenarios.

Preliminarily, documents were subjected to the following pre-processing steps: (1) First, we removed all words occurring in a list of common stopwords, as well as punctuation marks and numbers; (2) then, we extracted all n-grams, defined as sequences of maximum three words consecutively occurring within a document (after stopword removal)5; (3) at this point we have randomly split the set of seen data into a training set (70%), on which to run the GA, and a validation set (30%), on which tuning the model parameters. We performed the split in such a way that each category was proportionally represented in both sets (stratified holdout).

Based on the term frequency and inverse document frequency, the term weight will be calculated.

Term Weight = Term Frequency * Inverse Document Frequency

$$\text{Term Frequency} = \frac{\text{Term Count}}{\text{Total number of documents in count.}}$$

$$\text{Inverse Document Frequency} = \frac{\text{Total no. of documents}}{\text{No. of documents in the term}}$$

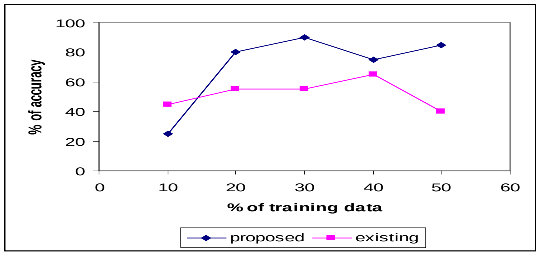


Fig4. Accuracy Vs Training Data

The above figure shows that the proposed technique gives better performance. Researchers showed 68% accuracy using the existing method with 31% test data while the technique is better both in accuracy and percentage of test data. Moreover it required processing for each class during training. But the proposed Algorithm does not require such process during training phase and hence reduces time.

6. Conclusion

The cluster based niche memetic and genetic algorithm have been designed and implemented by optimizing feature selection of the text in the documents repository. The efficacy of niche memetic is in evaluating optimal feature selection of text from the given set of documents. The contribution of genetic algorithm works on the evaluation of fitness function to cluster the relevant feature selected text which categorize documents to its most relevant cluster. In genetic algorithm, parameter tuning plays an important role for optimal feature text selection. In our approach, a text is segmented into groups of syllables with various lengths. We should build an auto parameter tuning scheme based on text length not a rigid one. This will speed up the processing time a lot.

The hybrid algorithm efficiency is shown in the experimental results, which confirms simultaneous global clustering and feature subset optimization mechanism is effective in text document categorization. The resulting algorithm is generally able to select relevant features and locate appropriate partitioning with the correct number of clusters and outperforms other methods implemented for comparison. We presented a new hybrid technique for text document clustering. The existing algorithms require more data for training as well as the computational time of these algorithms also increases. In contrast to the existing algorithms, the proposed hybrid algorithm requires less training data and less computational time. In spite of the randomly chosen training set we achieved 78% accuracy for 50% training data. Though 85% accuracy was observed in 30% training data, a class could not be classified, so we dropped this position and increased training data set for more acceptable result.

7. References

[1] S. Areibi and Z. Yang, “Effective Memetic Algorithms for VLSI Design Automation = Genetic Algorithms + Local Search + MultiLevel Clustering,” Evolutionary Computation, vol. 12, no. 3, pp. 327- 353, 2004.

[2] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.

[3] S. Wu, A.W.C. Liew, H. Yan, and M. Yang, “Cluster Analysis of Gene Expression Database on Self-Splitting and Merging Competitive Learning,” IEEE Trans. Information Technology in Biomedicine, vol. 8, no. 1, 2004.

[4] Y. Zhao and G. Karypis, “Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering,” Machine Learning, vol. 55, no. 3, pp. 311-331, 2004

[5] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surverys, 2002, 34 (1): 1-47.

[6] H.K. Tsai, J.M. Yang, Y.F. Tsai, and C.Y. Kao, “An Evolutionary Approach for Gene Expression Patterns,” IEEE Trans. Information Technology in Biomedicine, vol. 8, no. 2, pp. 69-78, 2004.

[7] P. Baldi and G.W. Hatfield, DNA Microarrays and Gene Expression. Cambridge Univ. Press, 2002

[8] Xiaoli Li, Bing Liu, Learning to classify text using positive and unlabeled data. The International Joint Conference on Artifical Intelligence (IJCAI) (2003)

[9] W. Sheng, A. Tucker, and X. Liu, “Clustering with Niching Genetic K-Means Algorithm,” Proc. Genetic and Evolutionary Computation Conf. (GECCO ’04), pp. 162-173, 2004

[10] C. Wei, C.S. Yang, H.W. Hsiao, T.H. Cheng, Combining preference- and content-based approaches for improving document clustering effectiveness, Information Processing & Management 42 (2) (2006) 350–372.

[11] J. Kogan, C. Nicholas, and V. Volkovich, "Text Mining with Information-Theoretic Clustering," IEEE Computational Science and Eng., pp. 52-59, 2003

A. K. Santra received the P. G. degree and Doctorate degree from I.I.T., Kharagpur in the year 1975 and 1981 respectively. He has got 20 years of Teaching Experience and 19 years of Industrial (Research) Experience. His area of interest includes Artificial Intelligence, Neural Networks, Process Modeling, Optimization and Control. He has got to his credit (i) 35 Technical Research Papers which are published in National / International Journals and Seminars of repute, (ii) 20 Research Projects have been completed in varied application areas, (iii) 2 Copy Rights for Software Development have been obtained in the area of Artificial Neural Networks (ANN) and (iv) he is the contributor of the book entitled **"Mathematics and its Applications in Industry and Business"**, Narosa Publishing House, **New Delhi.** He is the recognized Supervisor for guiding Ph. D. / M. S. (By Research) Scholars of Anna University-Chennai, Anna University-Coimbatore, Bharathiyar University, Coimbatore and Mother Teresa University, Kodaikanal. Currently he is guiding 12 Ph. D. Research Scholars in the Department. He is a Life member of CSI and a Life member of ISTE.

C. Josephine Christy received her M.Sc., M.Phil., M.B.A.,from Bharathiar University, Coimbatore. Currently she is working as Asst.Professor in Bannari amman Institute of Technology, Sathyamangalam. Her area of interest includes Text Mining, Web Mining. She has presented 2 papers international conferences and 5 papers in national Conferences. She is a Life member of Computer Society of India and a Life member of Indian Society for Technical Education.

B.Nagarajan received his Ph. D. degree in Pattern Recognition in the year 2010 from Anna University, Chennai. He has been in the teaching profession for more than a decade since 1997. His areas of academic interest are Image Processing, Pattern Recognition and Neural Networks. He has worked as

Also, he has published 14 papers in International Journals and presented 18 research papers in the National/International Conferences. He is the reviewer / editorial board member of 10 international journals from various countries like Singapore, Hong Kong, Korea, United States, Thailand, Romania and India.