International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Linear Kernel with Weighted Least Square Regression Co-efficient for SVM Based Tamil Writer Identification

Thendral Tharmalingam, Vijaya Vijayakumar,

Abstract: ***Tamil writer identification is the task of identifying writer based on their Tamil handwriting. Our earlier work of this research based on SVM implementation with linear, polynomial and RBF kernel showed that linear kernel attains very low accuracy compared to other two kernels. But the observation shows that linear kernel performs faster than the other kernels and also it shows very less computational complexity. Hence, a modified linear kernel is proposed to enrich the performance of the linear kernel in recognizing the Tamil writer. Weighted least square parameter estimation method is used to estimate the weights for the dot products of the linear kernel. SVM implementation with modified linear kernel is carried out on different text images of handwriting at character, word and paragraph levels. Comparing the performance with linear kernel, the modified kernel with weighted least square parameter reported promising results.***

Index Terms: Weighted Least Square, Parameter estimation, Support vector machine, Tamil handwriting, Kernels, Modified Kernel.

I. INTRODUCTION

Writer identification (WI) is the research work carried out here with proposed modified linear kernel to identify the writers based on their Tamil handwriting. In the document images features are recognized based mainly on the pattern of the handwritings than the pattern of the images. Identifying and extracting features from Tamil writings become a more challenging task as the Tamil alphabets are more multifaceted in nature. When compared to western scripts and other Indian scripts, Tamil scripts exhibit a large number of classes, stroke order variation and two-dimensional nature [1]. Tamil language covers massive amount of character sets which has more challenging patterns like loops, crossing, junction, different directions and so on. In our previous works, writer identification task was formulated as multiclass classification problem and solved using supervised learning algorithm namely support vector machine. Support Vector machine (SVM) is an intelligent computing tool that is being successfully applied to a wide range of pattern recognition problems. SVM is centred on strong mathematical foundations and statistical learning theory but results in simple yet very powerful algorithms with high generalization power on unseen data. The work was carried out for three levels of text like character level text, word level text and paragraph level text by developing three independent datasets. A text document with 100 paragraphs (20 pages) was designed and prepared by considering the intricacies involved in Tamil Alphabets/characters. These text dependent documents written by 300 individuals were collected and converted into digital images. The paragraph text image was segmented into words, further segmented into characters and 100 words were chosen at random for each writer. The word and character level text images were pre-processed using image processing tasks such as binarization, edge detection and thinning [2]. As the writing pattern of the same individual may vary at different instance, both global features and local features have been extracted from handwritten text images to make writer prediction more accurate. Global features are features taken by considering the text image as image rather than handwriting. These features were extracted from texture of the image using Gabor filters and co-occurrence matrices. Local features are features taken by considering the text image as handwriting. Various structural properties of the handwriting can be derived as local features. The local features were grouped into word measurement features, morphological features, fractal features, GSC features comprising gradient structural, concavity attributes. Word measurements features such as area, length, height, height from baseline to upper edge, height from baseline to lower edge, ascender line, and descender line were considered. Slope angle, junctions, loops are geometrical features, morphological features like directional opening, directional closing, directional erosion, k-curvature, fractal features like skew angle, slant angle, height of three main handwriting zones (upper zone, lower zone and middle zone), average width of writing have also been considered [3]. Two distinct datasets for character level and word level training were developed. The third training dataset was developed based on paragraph level text [4] using the same Tamil handwritten text corpus. As the first step in pre-processing, normalization was carried out in order to correct the skewed words in the handwriting image [5].

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Thendral Tharmalingam*, Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India.
Vijaya Vijayakumar, Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

The space between vertical and horizontal lines has been normalized using the horizontal projection profile method to produce a well-defined pattern for texture analysis. The images were converted into grayscale images to carry further pre-processing tasks. The pre-processing tasks such as edge detection, image dilation and box bounding were carried out to extract highly discriminative features.

The distinct features such as Gabor Filter [6], Gray Level Co-Occurrence Matrix (GLCM) [7], Generalized Gaussian Density (GGD) [8], Contourlet GGD [9], and directional features [10] were computed from pre-processed images. The particle swam optimization feature selection method was used to select the well contributing features to create the dataset.The support vector machine-based models were learned using the normalized training datasets with linear, polynomial and RBF kernels for multi class classification and the classifiers were built by tuning the SVM parameters for different levels of writer identification. The cross-validation results of SVM based models were analyzed and reported in Table I given below.

Table I. Results of our Previous Contribution

Parameters	Accuracy		
Levels of WI	**Lin**	**Polynomial**	**RBF**
Character	70.6	89.2	90.6
Word	75	91.2	93.8
Paragraph	88	93.6	96.2

From the observations it is clearly understood that linear kernel attains very less accuracy compared to polynomial and RBF kernels. It is also evidenced from the literature that in general most of the SVM implementations proves that RBF kernel is superior. But linear kernel is worthy when there is more number of features and best suitable for any pattern recognition task. Also computational complexity is very less in linear kernel and performs training faster. Hence it is proposed to employ the new form of linear kernel by considering the accuracy requirement, computational time, computational complexity and the nature of the problem. Weighted least square (WLS) parameter estimation is used to estimate the weight co-efficient of the linear kernel. This new form of linear kernel with distinctive properties will allow SVM algorithm to find better optimal hyperplane that discriminates writers in the feature space.

II. LINEAR KERNEL WITH WLS

The L2 norm SVM formulation is given by,

$$\min_{w,\gamma,\xi} \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{m} \xi_i^2$$

Subject to

$$d_i(w^T x_i - \gamma) + \xi_i - 1 \geq 0,\ 1 \leq i \leq m$$

$$\xi_i \geq 0,\ 1 \leq i \leq m$$

The lagrangian of the objective function [12] is,

$$L(w, \gamma, \xi, u) = \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{m} \xi_i^2 - \sum_{i=1}^{m} u_i [d_i(w^T x_i - \gamma) + \xi_i - 1]$$

$$= \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{m} \xi_i^2 - (\sum_{i=1}^{m} u_i d_i x_i^T) w - (\sum_{i=1}^{m} u_i d_i) \gamma - \sum_{i=1}^{m} u_i \xi_i + \sum_i u_i \quad (1)$$

Where u are the lagrangian multipliers.

Solving this with lagrangian duality based on parameters w, γ and ξ the dual problem is obtained as,

$$\max_u L(u) = \sum_{i=1}^{m} u_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} d_i d_j u_i u_j (x_i^T x_j + \frac{\delta_{ij}}{C})$$

Subject to

$$\sum_{i=1}^{m} d_i u_i = 0 \quad (2)$$

$$u_i \geq 0 \qquad 1 \leq i \leq m$$

The standard form in matrix format is,

$$\min_u L(u) = \frac{1}{2} u^T D (A A^T + \frac{1}{C}) D u - e^T u \quad (3)$$

Subject to

$$d^T u = 0$$

$$u \geq 0$$

(or)

$$\min_u L(u) = \frac{1}{2} u^T Q u - e^T u \quad (4)$$

Subject to

$$d^T u = 0$$

$$u \geq 0$$

Q can be computed as $Q = (A * A^T + I/C) .* (d * d^T)$ (5)

It is noted that the algorithm SVM finally requires three pieces of data Q, d and C where C is the regularization parameter, d is the diagonal matrix of class labels.

Q is the obtained from A * AT and d * dT [12].

$$AA^T = \begin{pmatrix} x_1^T x_1 & x_1^T x_2 & . & x_l^T x_m \\ . & . & x_i^T x_j & . \\ . & . & . & . \\ x_m^T x_1 & x_m^T x_2 & . & x_m^T x_m \end{pmatrix} = K \quad (6)$$

The i , jth element of A^{A^T} is $x_i^T x_j$ i.e. a dot product of two feature vectors xi and xj. The matrix K is called the linear kernel matrix which implies that all information needed for training is captured in the form of dot products of the training vectors. K is positive definite matrix and the set of kernels satisfy closure property. Complex kernels can be defined using simple one and employed in SVM for better learning. Some of the forms of linear kernel K are stated below,

1. K1 = A^{A^T} (7)
2. K2 = a. A^{A^T} (8)
3. K3 = $A^{A^T} + b$ (9)
4. K4 = a $A^{A^T} + b$ (10)

In this work the linear kernel K2 is used and the parameter a is obtained using parameter estimation method such as weighted least square regression. The constant vector a is of dimension equal to number of samples in the training dataset and each element is a weight added to the sum of the squares of the features.

A. *Weighted Least Square Regression*

In weighted least squares the distribution of the errors is unknown and permits general forms of unknown heteroscedasticity [14]. Weighted least squares reflect the behavior of the random errors in the model and it can be used with functions that are either linear or non-linear in the parameters. It works by adding extra weights to each data points, into the fitting criterion. The size of the weight identifies the exactness of the information contained in the associated observation. Enhancing the weighted fitting criterion to find the parameter estimates allows the weights to determine the influence of each observation to the final parameter estimates. So weight of each observation is given relative to the weights of the other observations. It is proved that different sets of absolute weights can have equal effects. Weights can be calculated using the following methods [15],

1. Weighted least squares is used when the variance is non-constant.
2. If the variance of the ith observation is σ_i^2, then weights $w_i = \frac{1}{\sigma_i^2}$ are used to determine the standard errors of coefficients, with $w_i = \frac{1}{\sigma_i^2}$, $\sum w_i e_i^2 = \sum \frac{1}{\sigma_i^2} e_i^2 = \sum (\frac{e^i}{\sigma^i})^2$ (11)
3. The sum of squares of the standardized errors is minimized to obtain the parameter estimates when weights = 1/variance. (12)

III. EXPERIMENT AND RESULTS

Tamil writer identification model is built to overcome the number of challenges involved in recognizing the writers of Tamil alphabets. Experiments are carried out by varying the datasets in different levels like character, word and paragraph. The model is developed using supervised learning algorithm like support vector machine. Support vector machine based model is trained using the training instances and the model identifies the test labels indicating which class the instance belong to. The developed model plays a major role to enrich the writer identification with an improved performance of the classifier. Support vector machine (SVM) is known for its kernel methods which consist of class of algorithms for pattern classification. A kernel method uses kernel functions, which enable them to operate in a high dimensional dataset. Generally in SVM based experiments three commonly used kernels are linear, polynomial and RBF. In all our previous experiments, RBF kernel based models showed better performance than linear and polynomial kernels. In this work linear kernel is a new form of linear kernel is defined with the aim of improving the performance of the SVM classifier by adding weights into the dot products of the linear kernel. The weights for the weighted linear kernel are estimated using weighted least square parameter estimation method. SVM with weighted linear kernel is implemented for three independent datasets and is compared with the linear kernel to evaluate the writer identification model. The performance of weighted linear kernel is analyzed with different levels of datasets character (TWINC), word (TWINW) and paragraph (TWINP). The datasets are partitioned into 80% training text images to model the data and 20% of the testing text images to predict the writer. Every level of dataset contains 30000 images of Tamil handwriting with multiclass labels 1 to 300. The profile of dataset is shown in Table II. The predictive accuracy (PA), precision, recall and F-measure are observed for the trained models. Prediction accuracy is the ratio of number of correctly classified instances and the total number of instances. Precision is the segment of retrieved instances that are relevant, recall is the fraction of relevant instances that are retrieved and F-measure computes the average of the information retrieval in precision and recall. SVM with WLK kernel has been implemented for three datasets by tuning C-regularization parameter and the predictive accuracies of classifiers are shown in Table III. Table IV and Table V shows the performance analysis of proposed weighted linear kernel with various levels of features. The performance comparison of the modified weighted linear kernel with linear (lin) kernel is shown in Table VI.

Table II. The profile of dataset

Levels of dataset	Character	Word	Paragraph
Number of data	30000	30000	30000
Training data	24000	24000	24000
Testing data	6000	6000	6000

Number of features	26	422	422
Number of Class labels	1-300	1-300	1-300

Table III. SVM with WLK kernel by tuning C- regularization parameter

Levels of WI	**Character**			**Word**			**Paragraph**		
Para meters	C=1	C=5	C=10	C=1	C=5	C=10	C=1	C=5	C=10
WLK-SVM	72.3	70.6	71.4	75	77.2	72	79.2	80.4	90.8

Table IV. Performance analysis of proposed weighted linear kernel with various levels of features

Parameters	**Class**	**Training Images**	**Testing Images**	**Accuracy (%)**	**Precision**	**Recall**	**F-measure**
Character level features	1	24000	6000	72.9	0.65	0.68	0.66
	2	24000	6000	70.5	0.61	0.67	0.64
	3	24000	6000	73.8	0.66	0.69	0.67
	4	24000	6000	74	0.66	0.69	0.67
	5	24000	6000	74.2	0.66	0.69	0.67
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	296	24000	6000	70.8	0.62	0.67	0.65
	297	24000	6000	69.8	0.60	0.67	0.63
	298	24000	6000	71.5	0.63	0.67	0.65
	299	24000	6000	72.6	0.64	0.68	0.66
	300	24000	6000	73.1	0.65	0.69	0.67
Word level features	1	24000	6000	77.89	0.71	0.71	0.71
	2	24000	6000	75	0.68	0.70	0.69
	3	24000	6000	76.1	0.69	0.70	0.69
	4	24000	6000	78.2	0.72	0.72	0.72
	5	24000	6000	79	0.73	0.72	0.73
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	296	24000	6000	78.7	0.72	0.71	0.71
	297	24000	6000	76.9	0.72	0.76	0.74
	298	24000	6000	77.9	0.71	0.70	0.70
	299	24000	6000	75	0.68	0.70	0.69
	300	24000	6000	77.7	0.72	0.73	0.73
Paragraph level features	1	24000	6000	90.7	0.88	0.77	0.82
	2	24000	6000	89.8	0.87	0.77	0.82
	3	24000	6000	87.9	0.85	0.76	0.80
	4	24000	6000	93	0.91	0.76	0.83
	5	24000	6000	93.7	0.92	0.76	0.83
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	296	24000	6000	94.9	0.93	0.75	0.83
	297	24000	6000	91.8	0.89	0.76	0.82
	298	24000	6000	88.8	0.86	0.77	0.81
	299	24000	6000	87.9	0.85	0.76	0.80
	300	24000	6000	89.9	0.87	0.75	0.81

Table V. Overall performance of the WLK-SVM Kernel

Parameters	Accuracy (%)	Precision	Recall	F-measure
Character	72.3	0.733	0.968	0.834
Word	77.2	0.771	0.889	0.826
Paragraph	90.8	0.915	0.8318	0.871

Table VI. Relative Measurements of Linear and WLK-SVM Kernel

Parameters	Accuracy (%)		Precision		Recall		F-measure	
Levels of Writer Identification	Lin	WLK-SVM	Lin	WLK-SVM	Lin	WLK-SVM	Lin	WLK-SVM
Character	70.6	72.3	0.689	0.733	0.827	0.968	0.751	0.834
Word	75	77.2	0.706	0.71	0.748	0.889	0.726	0.826
Paragraph	88	90.8	0.942	0.915	0.989	0.8318	0.964	0.871

Receiver Operating Characteristic (ROC) curves for ten output classes are plotted. The more each curve squeezes the left and top edges of the plot, the better the classification. ROC based on precision and recall in character level, word level and paragraph level are depicted in Fig. 1. to Fig. 3.

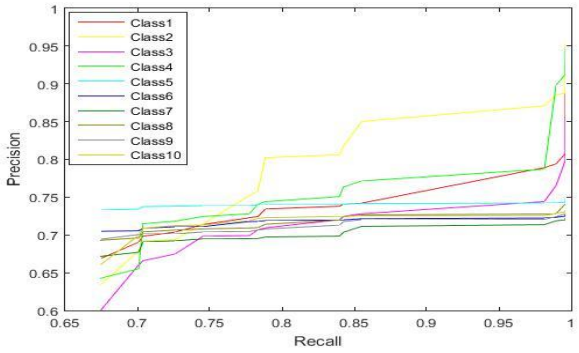

Fig. 1. ROC for character level dataset

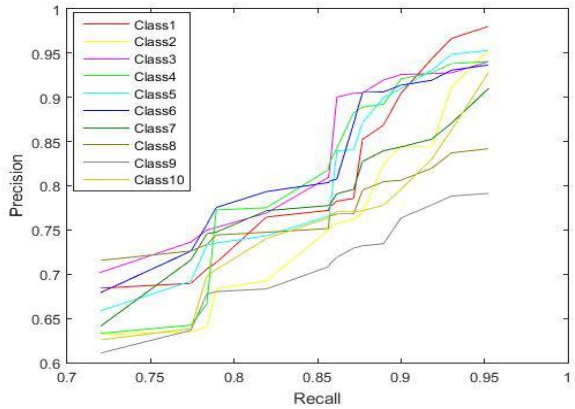

Fig. 2. ROC for word level dataset

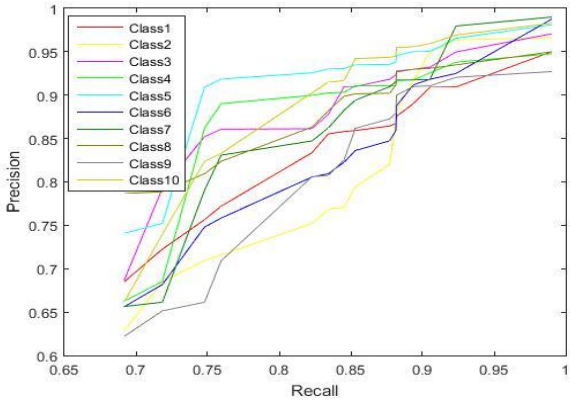

Fig. 3. ROC for paragraph level dataset

It is observed that ROC curve for character level, class 4 have high precision of 0.95 and class 3 is curved low at 0.60. In word level, class 1 have high precision of 0.97 and class 9 is curved low at 0.62. In paragraph level, class 7 have high precision of 0.99 and class 9 is curved low at 0.64. The performance of the weighted linear kernel based SVM prediction models is observed in terms of accuracy for all three datasets and is illustrated in Fig. 4. The comparative performance of the weighted linear kernel (WLK) over linear kernel for all three levels of writer identification is illustrated in Fig. 5.

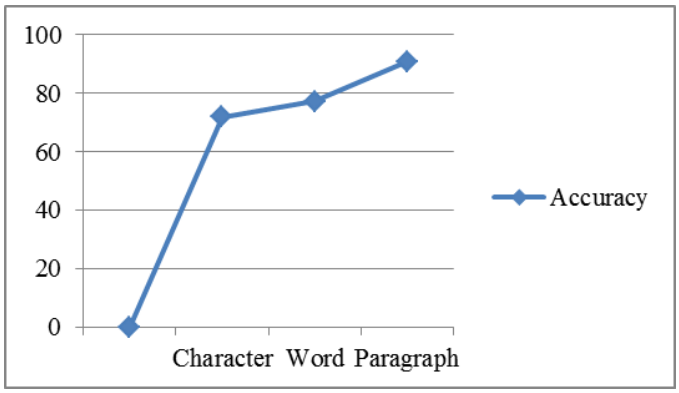

Fig. 4. Prediction Accuracy of WLK

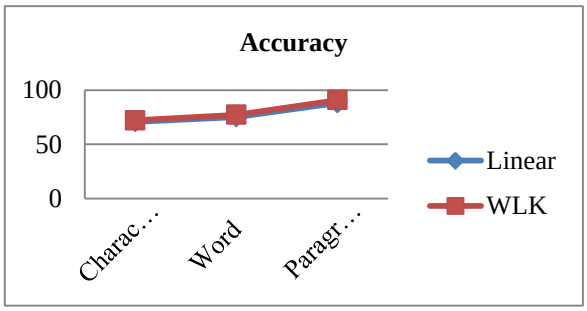

Fig. 5. Prediction Accuracy of WLK and linear kernel

A. *Findings*

Linear kernel is a dot product of features with less computational complexity.

The system optimizes only C regularization parameter in linear kernel which makes it faster than other kernels. Hence the modified weighted linear kernel determines the optimum hyperplane with less computational complexity and achieved better performance. Due to less computational complexity, the time taken to train the model is very less. From the above qualified analysis it is observed that the modified weighted linear kernel based prediction model shows (90.8%) in paragraph level, 77.2% in word level and 72% in character level which confirms high accuracy than in linear kernel which confirms (88%) in paragraph level, 75% in word level and 70.6% in character level. The prominent accuracy of 90.8% is achieved using the modified weighted linear kernel in paragraph level.

IV. CONCLUSION

In this work writer identification is focused on Tamil language using a modified form of linear kernel in SVM. The novelty is introduced in linear kernel and the framework of linear kernel based method is modified using parameter estimation technique. The proposed weighted linear kernel (WLK-SVM) is developed using weighted least square parameter estimation technique and the writer identification models were built. The models were built using 24000 training images and 6000 testing images. The proposed weighted linear kernel performs competitively better in comparison with the widely used polynomial and RBF kernels. The use of WLK-SVM is more beneficial in terms of less computational complexity, minimum time, scalability yielding better results than the linear kernel based SVM method. This work can be extended in future with different parameter estimation techniques.

REFERENCES

1. Antony Robert Raj, Dr.S.Abirami, “A Survey on Tamil Handwritten Character Recognition using OCR Techniques, CCSEA, SEA, CLOUD, DKMP”, CS & IT 05, DOI : 10.5121/csit.2012.2213, pp. 115–127, 2012.
2. T. Thendral, M.S.Vijaya, ”Analysis of Tamil Character Writings and Identification of Writer Using Support Vector Machine”, IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 8-10 May, 2014.
3. T. Thendral, “Discovering Tamil Writer Identity Using Global and Local Features of Offline Handwritten Text”, International Review on Computers and Software (IRECOS), Vol. 8 No 9, Page 2080 – 2087, 2013.
4. T. Thendral, ”Prediction of Writer Using Tamil Handwritten Document Image Based on Pooled Features”, World Academy of Science, Engineering and Technology, Vol. 9, No. 6, Page 1481 – 1487, 2015.
5. H. E. S. Said, G. S. Peake, T. N. Tan and K. D. Baker, “Writer Identification from Non-uniformly Skewed Handwriting Images”, British Machine Vision Conference.
6. D. M. Tsai, “Optimal Gabor filter design for texture segmentation”, Image and Vision Computing, 2001, Volume (19), Page: 299—316.
7. Revathi S V and Vijaya M S, “Predicting the Identity of a Person using Aggregated Features of Handwriting”, Advances in Image and Video Processing, Volume 2 No 6, Dec (2014); pp: 23-33
8. Zhenyu Heb, Xinge Youb, Yuan Yan Tanga, “Writer identification using global wavelet-based features”, Elsevier publication, Neurocomputing (2008), page: 1832–1841.
9. Z.Y.Tang. & X.You (2005). A Contourlet based method for writer identification, In Systems, Man and Cybernetics, 2005, IEEE International Conference on Volume 1, pp. 364-368. IEEE.
10. Marius Bulacu, Lambert Schomaker, Louis Vuurpijl. (2003), “Writer Identification Using Edge-Based Directional Features”, Proceedings of the Seventh International Conference on Document Analysis and Recognition.
11. P. B. Pati, A. G. Ramakrishnan, OCR in indian scripts: A survey, IETE Technical Review, 22(3), May-June 2005, 217-227.
12. K.P. Soman, R. Loganathan, V. Ajay, Machine Learning with SVM and Other Kernel Methods, PHI Learning Pvt. Ltd., 02-Feb-2009.
13. Yong Wang, Hui Guo, Weld Defect Detection of X-ray Images Based on Support Vector Machine, Journal IETE Technical Review, 03 Jun 2014, Pages 137-142.
14. Shakeeb Khan, Arthur Lewbel, Weighted and Two Stage Least Squares Estimation of Semiparametric Truncated Regression Models, Econometric Theory, 23, 2007, 309–347, May 2003.
15. Sulaimon Mutiu O., Application of Weighted Least Squares Regression in Forecasting, International Journal of Recent Research in
16. Interdisciplinary Sciences (IJRRIS), Vol. 2, Issue 3, pp: (45-54), July 2015 - September 2015.
17. Surendra Kumar Rakse, Sanyam Shukla, Spam Classification using new kernel function in Support Vector Machine, International Journal on Computer Science and Engineering, Vol. 02, No. 05, 2010, 1819-1823.
18. R. Sangeetha, B. Kalpana, Performance Evaluation of Kernels in Multiclass Support Vector Machines, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-5, November 2011.
19. N.E. Ayat, M. Cheriet, L. Remaki, C.Y. Suen, KMOD-A New Support Vector Machine Kernel With Moderate Decreasing for Pattern Recognition Application to Digit Image Recognition.

AUTHORS PROFILE

Mrs.T. Thendral has nine years of teaching experience. She is pursuing her doctoral program in the area of Data Mining. She has worked as a member in various Seminars and conferences. She has participated in various Conferences, Seminars, Faculty Development Programme and Workshops. Her areas of interest include Data Mining, Pattern Recognition, Support Vector Machine and Machine learning.

Dr.M.S.Vijaya has 30 years of teaching experience and 15 years of research experience. She has completed her doctoral programme in the area of Natural Language Processing. She is a member of Computer Society of India, Advanced Computing and Communications Society (IISc, Bangalore), International Association of Engineers (Hong Kong), International Association of Computer Science and Information Technology (IACSIT, Singapore). She has served as a member of programme committee in several International conferences. Her areas of interest include Data Mining, Support Vector Machine, Machine learning, Pattern Recognition, Natural Language Processing and Optimization Techniques. She has also completed various funded research projects.