Machine Learning Based Supervised Feature Selection Algorithm for Data Mining

K.Sutha, J. Jebamalar Tamilselvi

Abstract: ***Data Scientists focus on high dimensional data to predict and reveal some interesting patterns as well as most useful information to the modern world. Feature Selection is a preprocessing technique which improves the accuracy and efficiency of mining algorithms. There exist a numerous feature selection algorithms. Most of the algorithms failed to give better mining results as the scale increases. In this paper, feature selection for supervised algorithms in data mining are considered and given an overview of existing machine learning algorithm for supervised feature selection. This paper introduces an enhanced supervised feature selection algorithm which selects the best feature subset by eliminating irrelevant features using distance correlation and redundant features using symmetric uncertainty. The experimental results show that the proposed algorithm provides better classification accuracy and selects minimum number of features.***

Index Terms: Feature Selection, Supervised learning, Data mining

I. INTRODUCTION

Data Analysts use data mining tools in various fields to improve the quality of service. Some of the examples are listed: Analysis of vast amount of patient health data helps Healthcare providers to improve the quality of treatment, patient satisfaction and patient-centric care. Customer-centric companies analyze their massive amount of customer data to enhance customer satisfaction, to understand the customers purchasing habit and to take better decision to improve their profitability. Banking sectors analyze their large amount of customer data to provide better financial services and to safeguard them from fraudulent customers.

Data with high dimensionality not only posses useful information but also have some irrelevant, noise and redundant data. Applying mining techniques on massive data without preprocessing doesn't provide accurate results and it also affects the performance of mining algorithms. Drastically and continuously growing size of dataset keeps Feature selection as the active research topic for many years, even today. Feature selection (FS) aims at finding an optimal feature subset from original feature set, by removing features which are out of point of interest. It results in lowering computational cost, improving the effectiveness and efficiency of mining performance and increasing comprehensibility [1]. The optimal feature subset provides better model construction during classification.

The dataset may be labeled, partially labeled or unlabeled. Depending upon the presence or absence of class information, the feature selection algorithms are categorized as supervised, unsupervised and semi-supervised [2][3][4]. Supervised learning deals with labeled data.

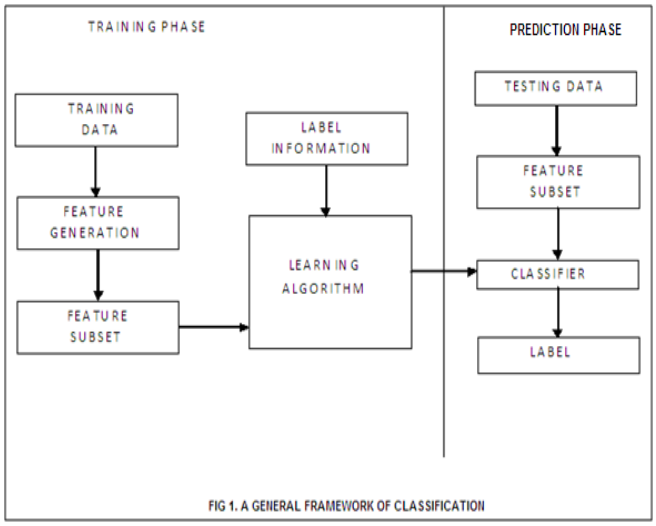

FIG 1. A GENERAL FRAMEWORK OF CLASSIFICATION

Fig 1 shows the classification process [1] which includes the feature selection process. The entire dataset is partitioned into training and testing dataset. During training phase, the Feature Generation module chooses a subset of features. The feature subset along with label information is given to the supervised learning algorithm. The learning algorithm constructs a model or classifier, on the basis of training data and label information provided. In the prediction phase, the classifier obtained from the training phase, is provided with a new set of samples from testing dataset with selected features, results in label information. The classification performance determines the goodness of the selected feature subset.

Some of the supervised feature selection algorithms are presented in the next section.

II. EXISTING SUPERVISED FEATURE SELECTION ALGORITHMS

In this section some of the existing supervised feature selection algorithms are discussed. Recursive Feature Elimination (RFE) eliminates non-discriminative features iteratively. It is widely used in binary classification [5].

Revised Manuscript Received on August 05, 2019

K.Sutha Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India.

Dr.J. Jebamalar Tamilselvi, Professor, Department of MCA, Jaya Engineering College, Chennai, Tamil Nadu, India.

RFE is a supervised learning technique, it operates as follows [6][7],It first trains the classifier with original feature set and calculates the weight for each feature in the feature set with certain criterion on the basis of their involvement in classification. Features with lowest weight are then eliminated. This process is repeated until optimal feature subset is obtained. Drawback is that the recursive iterations of RFE make it computationally expensive [7]. Selection of F number of top features, out of T number of total features, requires T-F iterations by eliminating only one lowest weight feature at a time. As the number of features elimination for each iteration increases, the performance degrades. Random Forest [8] is an improvement of Decision trees concept and plays a better role in feature selection. It works by ranking features using Gini Impurity or Information gain. Gini Impurity is the measure of incorrect classification of a new sample. It is calculated as,

$$G(k) = \sum_{i=1}^{j} P(i) * (1 - P(i))$$

Where P(i) , the probability of certain classification. Random tree provides better predictive accuracy. It reduces the computation cost and also controls the problem of over-fitting.

Univariate Feature Selection [10] is a fast and efficient feature selection technique which falls under Filter model. It is most widely used in microarray studies [10]. Selection of best feature subset is done on the basis of feature weights assigned to each feature by performing some univariate statistical tests [11] such as chi-square, Fisher Score, Information Gain etc. The Chi-Square[14] is a statistical method of determining the correlation between the classes and the features. It is zero if both the class and feature are independent. Chi square score for NC (no. of classes) and a feature with nf (no. of different values) is calculated as,

$$\chi^2 = \sum_{i=1}^{nf}\sum_{j=1}^{NC}\frac{(n_{ij} - X_{ij})^2}{X_{ij}}$$

Where n_{ij} is the number of samples in the i^{th} feature and j^{th} class . X_{ij} is the expected frequency of n_{ij} . It is calculated as $X_{ij} = \frac{Q*\sum_{i=1}^{Nf} n_{ij}}{N}$ where, N= $\sum_{i=1}^{Nf} Q$ and Q=$\sum_{j=1}^{NC} n_{ij}$. Information Gain (IG) or Mutual Information [18] measures how much a feature(X) is relevant to the class(Y). IG(X|Y) = E(X) – E(X|Y), E(X) is the entropy of X and E(X|Y) is the entropy of X after observing Y. Entropy(E) measures the uncertainty associated with a random variable X. Entropy, $E(X) = -\sum_{x\in X} p(x)log_2(p(x))$ and $E(X|Y) = -\sum_{y\in Y} p(y)\sum_{x\in X} p(x|y)log_2(p(x|y))$. Fisher Score [12][16] is a filter type of supervised feature selection technique, measures the significance of each feature by comparing it's correlation to the output labels. The Fisher score for a feature is calculated as,

$$FS(i) = \left| \frac{mp_i - mn_i}{(sp_i)^2 - (sn_i)^2} \right|$$

Where mp_i - mean of feature in the positive class , mn_i - mean of feature in the negative class, sp_i - standard deviation feature in the positive class and sn_i - standard deviation feature in the negative class. Once all the features are ranked, the specific number of features with maximum score is selected as feature subset.The major drawback of Univariate feature selection is that it cannot handle redundancy.

Support Vector Machine is a classification method, can be used in feature selection process [12], which falls under wrapper model. Wrapper methods are computationally expensive but give accurate classification results comparing with the filter ones [13]. I.Guyon et al.[5] used the SVM classifier in feature ranking and the recursive feature elimination technique to remove the low ranked features to select the best feature subset. The weights of SVM classifier is used in ranking the features [17]. SVM as a classifier offers higher classification accuracy compared with other classifiers. It is rarely used because of its higher learning time in case of larger datasets [15]. Therefore it is also necessary to provide SVM with optimal feature subset to reduce the computation time of classification process.

Hall (2000) measures the goodness of feature subset, based on the hypothesis that "features in the best feature subset are strongly correlated to the target class, yet uncorrelated to each other". Following this, FCBF [20] measures predominant correlation using Symmetric Uncertainty (SU), to select the best feature subset. Symmetric Uncertainty [19] measures the predominant nature of features in the classification process. It is defined as,

$$SU = 2\left[\frac{I(X|Y)}{E(X)+E(Y)}\right]$$

Where I(X,Y) , E(X) and E(X|Y) are defined in section 3.3. The FCBF algorithm ranks features using SU coefficient, find out class-feature relevance and selects the relevant features based on some predefined threshold value. The selection process stops when the feature subset has *nlogn* number of features. Relevant features are sorted in descending order according to their SU value. Symmetry is a preferred property in measuring correlation between features. In the second part, SU measures the correlation between features. The predominant features are selected as the best feature subset. FCBF handles both irrelevant and redundant features.

Szekely et al. introduced Distance Correlation [21] to measure the linear and non –linear relationship between two variables. Pearson's Correlation can measure only linear dependency. The squared distance correlation [21][22] is defined as

$$D^2(x,y) = \begin{cases} \dfrac{V^2(x,y)}{\sqrt{V^2(x,x)V^2(y,y)}}, & V^2(x,x)V^2(y,y) > 0. \\ 0, & V^2(x,x)V^2(y,y) = 0. \end{cases}$$

Where $D^2(x,y)$ is equal to 0 if and only if both the variables are independent and $D^2(x,y)$ satisfies the relation $0 \leq D^2(x,y) \leq 1$. $V^2(x,y)$ is the squared distance covariance[21][22] , defined as the weighted L2 distance between fX,Y (t, s) and fX(t) · fY (s),

$$V^2(x,y) = \int_{\mathbb{R}^{m+n}} |f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2 w(t,s)\,dt\,ds$$

$w(t, s) = (C_p C_q |t|_p^{1+p} |s|_q^{1+q})^{-1}$,C_p and C_q are constants

where $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$, f_X **and** f_Y are marginal characteristic functions of **x and y**, $f_{X,Y}$ is the joint characteristic function, and $w(t, s)$ is a weight function .Minimum Redundancy Maximum Relevance [23] is a filter model, which selects relevant features using mutual information as a measure. The maximum mutually different features are selected to attain minimum redundancy. The minimum redundancy condition is given by Min WI, WI = $\frac{1}{|S|^2}\sum_{i,j \in S} I(i,j)$ where S is the set of features, I(i,j) is the mutual information between features I and j. The Maximum relevance condition is given by Max VI, VI= $\frac{1}{|S|}\sum_{i \in S} I(C,i)$, where C is the target class, I(x, y) is the mutual information of variables x and y, I(x, y)= $\sum_{i,j} p(x_i,y_j) \log \frac{p(x_i,y_j)}{p(x_i)p(y_j)}$

ReliefF [24] is a filter model supervised feature selection algorithm. It can handle incomplete, noisy and multi-class datasets. ReliefF is an efficient and reliable algorithm but it failed to handle redundancy [25].

PROPOSED ALGORITHM

Most of the feature selection algorithms consider only the relevant features during features selection process. The presence of redundant features affects the performance of classification process. An effective and efficient feature selection algorithm should take into account of eliminating redundant features to get the optimal feature subset. This paper introduces a new algorithm which involves in the selection of optimal feature subset in addition to it also handles redundancy.

Algorithm: D-Sym

Input : DS($F_1, F_2, F_3, \ldots\ldots C$) // Training Set

λ // threshold value

Output: $Optimal_{List}$ // Best feature subset

// **Relevance analysis :** Calculate DCor(X,C) and create an ordered list Temp_{List} of features .

1. begin
2. for i = 1 to M do
3. compute $\text{DCor}_{i,c}$ for all features F_i
4. if ($\text{DCor}_{i,c} >= \lambda$)
5. Add F_i to Temp_{List}
6. end
7. Sort Temp_{List} in descending order of $\text{DCor}_{i,c}$ value

//Redundancy analysis :

// Take the first feature from the Temp_{List} .Compute and compare SU for the first and the remaining features in the list .

//Remove all features for which SU is greater than SU with //Class C.

8. F_a = assignFirstElement (Temp_{List})
9. do
10. F_b = assignNextElement(Temp_{List} , F_a)
11. if (F_b ! = $NULL$)
12. do
13. $\text{Temp}_b = F_b$
14. if ($SU_{a,b} \geq SU_{b,c}$)
15. delete feature F_b from Temp_{List}
16. F_b = assignNextElement($Temp_{List}$, $Temp_b$)
17. else
18. F_b = assignNextElement($Temp_{List}$, F_b)
19. until(F_b == $NULL$)
20. F_a= assignNextElement(Temp_{List} , F_a)
21. until(F_a == $NULL$)
22. $Optimal_{List} = \text{Temp}_{List}$
23. end

The proposed algorithm works in two steps as shown in Fig2.

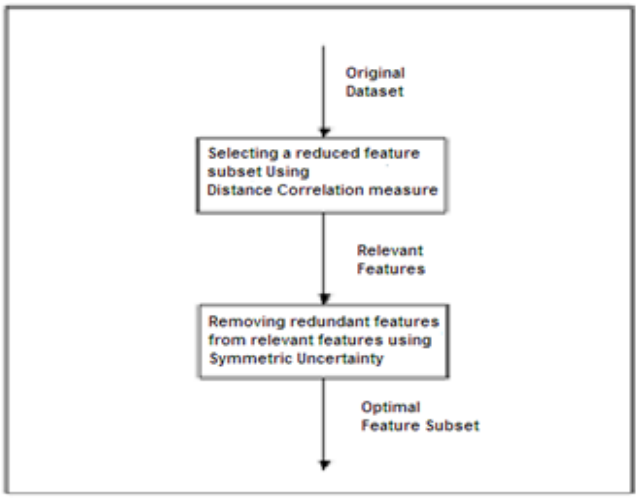

Fig 2. Proposed Model of Supervised Feature Selection

In the first part, distance correlation used in selecting relevant features. The squared distance correlation [21][22] is calculated as

$$D^2(x,y) = \begin{cases} \dfrac{V^2(x,y)}{\sqrt{V^2(x,x)V^2(y,y)}}, & V^2(x,x)V^2(y,y) > 0. \\ 0 & , V^2(x,x)V^2(y,y) = 0. \end{cases}$$

Where $D^2(x,y)$ is equal to 0 if and only if both the variables are independent

$D^2(x,y)$ satisfies the relation $0 \leq D^2(x,y) \leq 1$.

$V^2(x,y)$,the squared distance covariance[21][22] is calculated as

$$V^2(x,y) = \int_{\mathbb{R}^{m+n}} |f_{x,y}(t,s) - f_x(t) f_y(s)|^2 w(t,s)\,dt\,ds$$

$w(t,s) = (C_p C_q |t|_p^{1+p} |s|_q^{1+q})^{-1}$, C_p and C_q are constants

where $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$, f_x **and** f_y are marginal characteristic functions of **x and y**, $f_{x,y}$ is the joint characteristic function, and $w(t, s)$ is a weight function

The feature with highest distance correlation value is considered as the most important feature. Features with distance correlation greater than a predefined threshold value λ are selected as relevant features. The selected features are sorted in descending order of distance correlation value. The topmost N features are selected.

In the next part, Symmetric Uncertainty is applied to remove the redundant features **,** $SU = 2\left[\frac{I(X|Y)}{En(X)+En(Y)}\right]$. It is used to calculate pair-wise feature correlation between all the features and compared with the SU between feature and class. If $SU(F_i,F_j)$ is greater than $SU(F_j,C)$, the feature F_j is removed as an redundant feature. It is referred as predominant correlation [20] in FCBF algorithm. The remaining features form the best feature subset.

III. EMPIRICAL STUDY

A. *Dataset Description*

Datasets used in analyzing the performance of proposed algorithm and other algorithms are listed below. The number of features in the dataset ranges from 56 to 1301. The Classification accuracy of Knn Classifier is used to measure the goodness of the selected feature subset.

Table 1: Datasets

Dataset	No of Features	No of Instances	No of classes
Lungcancer	56	32	3
Hill Valley	101	1212	2
Mfeat	217	2000	10
ISVT	311	126	2
Isolet	618	7797	2
Micromass	1301	571	20

B. *Results and Analysis*

The performance of D-Sym algorithm is evaluated and compared with FCBF and MRMR. The number of features selected by the proposed algorithm and classification accuracy of the proposed algorithm are compared against FCBF and MRMR algorithms, shown in Table 2 and 3.

Table 2 : Comparison of no. of features selected by proposed method against other methods

Datasets	Proposed Algorithm	FCBF	MRMR
Lungcancer	**4**	4	8
Hill Valley	**6**	6	11
Mfeat	**25**	29	17
ISVT	14	10	**7**
Isolet	24	**23**	30
Micromass	**173**	---	227

Table 3 : Comparison of classification accuracy given by proposed method against other methods

Datasets	Proposed Algorithm	FCBF	MRMR
Lungcancer	**100**	66	75
Hill Valley	65.57	**68.03**	63.93
Mfeat	**96**	94	94
ISVT	92.8	**92.9**	71.42
Isolet	**73.17**	70.63	64.1
Micromass	**63.33**	---	53

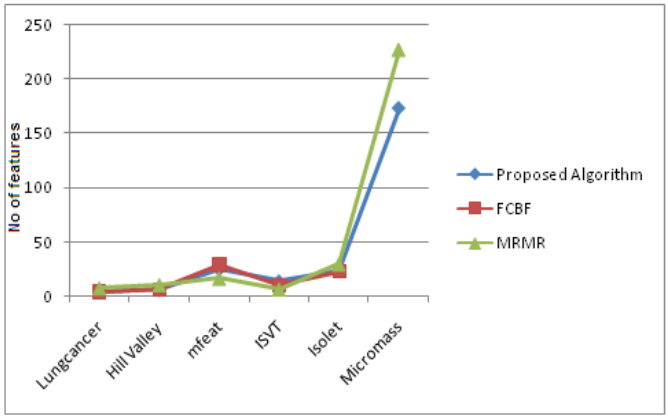

Fig. 3. Comparison Of No. Of Features Selected By Proposed And Other Methods

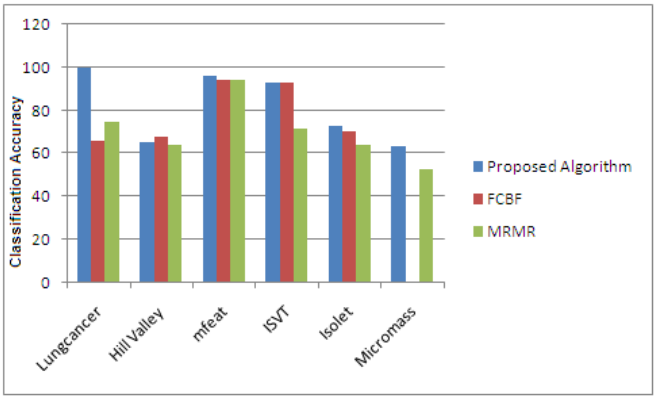

Fig.4 Classification Accuracy Of Proposed And Other Methods

Experimental results in Table 3 and Fig 4, reveals that the proposed algorithm achieves better classification accuracy when compared with FCBF and MRMR. FCBF failed to handle thousand numbers of features. The proposed algorithm successively scales up with the increasing number of features, FCBF cannot work with thousand number of features. Experimental results listed in Table 2 and Fig. 3, reveals that the proposed algorithm mostly selects least number of features when compared with FCBF and MRMR. In overall conclusion the proposed algorithm give better result as expected when compared with results of other existing algorithms.

IV. CONCLUSION

In this paper, a new supervised feature selection algorithm is proposed. It obtains an optimal feature subset by eliminating irrelevant features using distance correlation and redundant features are removed using symmetric uncertainty. Experimental result showed that the proposed algorithm scales up with the increasing dimension whereas FCBF failed to do so. It also selects an optimal feature subset with lesser number of features than MRMR. Thus the proposed algorithm is reasonably providing better performance as compared with FCBF and MRMR. The overall conclusion is that the proposed algorithm selects a least number of features with better classification accuracy. And in most cases, it outperforms the existing algorithms.

REFERENCES

1. J.Tang, S.Alelyani and Huan Liu, “Feature Selection for Classification: A Review”, Data Classification: Algorithms and Applications, Pages 37-64, 2014
2. J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping, “Use of the zero norm with linear
3. models and kernel methods”, Journal of Mach. Learning Research, 3:1439–1461, 2003.
4. J.G. Dy and C.E. Brodley. Feature selection for unsupervised learning. The Journal
5. of Mach. Learning Research, 5:845–889, 2004.
6. Z. Zhao and H. Liu, “Semi-supervised feature selection via spectral analysis”,
7. In Proceedings of SIAM International Conference on Data Mining, 2007
8. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik,” Gene selection for cancer classification using support vector machines”, Machine Learning, vol 46:389-422, 2002.
9. W You , Z Yang , G Ji,” Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination”, Pg No: 1463 - 1475
10. Hansheng Lei, Venu Govindaraju, “Speeding Up Multi-class SVM Evaluation by PCA and Feature Selection”, SIAM Int’l Conf. on Data Mining, Pg No: 72-79, April 23, 2005.
11. Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.
12. T A Alhaj , M Md Siraj,A Zainal, H T Elshoush,F Elhaj,”Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation”, 2016
13. Y.Sayes, I.Inza, P.Larranaga,” A Review of Feature Selection Techniques in Bioinformatics”, Bioinformatics, Vol 23,pp 2507-2517.
14. Z.Zhao, S.Sharma, F.Morstratter,S.Alelyani, A.Anand,H.Liu,”Advancing Feature Selection Research”,ASU Feature Selection Repository,2010
15. S Maldonado, R Weber,”A wrapper method for feature selection using Support Vector Machines”, Information Sciences, 179 (2009) 2208–2217
16. R.Kohavi , G.H. John, “Wrappers for Feature Subset Selection”, Artificial Intellegence, vol.97, 273-324,1997.
17. H. Liu and R. Setiono, “Chi2: Feature selection and discretization of numeric attributes”, pg 388-391, IEEE Computer Society.1995
18. S T Ikram , A K Cherukuri , “Intrusion detection model using fusion of chi-square feature selection and multi class SVM”, (2017) 29, pp 462–472
19. Chapman, Hall, “Data Classification Algorithms and Applications”, Data Mining and Knowledege Discovery Series.
20. D Roobaert, G Karakoulas, and N V. Chawla, “. Information Gain, Correlation and Support Vector Machines”, pp 463–470 (2006)
21. J Li, K Cheng, S Wang, F Morstatter, R P. Trevino, J Tang, H Liu,” Feature Selection: A Data Perspective”, Vol. 9, No. 4, 2010.
22. Ian H. Witten and E Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, II ed, 2005.
23. Lei Yu, Huan Liu,”Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution”, AZ 85287-5406, USA.
24. G. J. Szekely, M. L. Rizzo, N. K. Bakirov, "Measuring and testing independence by correlation of distances", Annals of Statistics,2007, 35 (6):Pg. No: 2769–2794.
25. Arin Chaudhuri and Wenhao Hu,” A fast algorithm for computing distance Correlation”,IOT, SAS Institute Inc.,2018
26. C. H. Q. Ding and H. Peng. “Minimum redundancy feature selection from microarray gene expression data”.. In CSB, pages 523-529. IEEE Computer Society, 2003.
27. I.Kononenko, “Estimating attributes : Analysis and extension of RELIEF”, pg 171-182, 1994
28. Q Song, J Ni, and G. Wang,” A Fast Clustering-based Feature Subset Selection Algorithm for High-Dimensional Data”, IEEE , Vol 25, No.1, Jan 2013.

AUTHORS PROFILE

K.Sutha is a research scholar at Bharathiar University, Coimbatore, Tamilnadu. She received her MCA degree from Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. Her area of interests includes Data Warehousing , Data Mining , and Big Data.

Dr. J. Jebamalar Tamilselvi received her Ph.D. in 2009 from the Department of Computer Applications at Karunya University, Coimbatore, INDIA. She received her B.Sc. (Computer Science) from Manonmanium Sundaranar University of Tamil Nadu, INDIA in 2003 and MCA Degree from Anna University, Coimbatore, Tamil Nadu, INDIA in 2006. Her area of interest includes Data cleansing approaches, Data Extraction, Data Integration, Data Warehousing and Data Mining. She is a life Member of International Association of Engineers (IAENG), International Association of Computer Science and Information Technology (IACSIT), and the Society of Digital Information and Wireless Communications. Reviewer and Member of International Journal of Engineering Science and Technology (IJEST) Member and Convergence Information Technology (JCIT). Her research has been accepted and published in 17 international journals, and 12 national and international conferences. She had been awarded the P.K Das Memorial Best Faculty Award in 2014 by the Nehru Group of Institutions, Coimbatore and the Education and Research Award in 2015 by the Karunya University, Coimbatore.