Available online at http://www.mecs-press.net/ijeme

A Comparative Study of Data Mining Algorithms for Image Classification

P Thamilselvan[a], Dr. J. G. R. Sathiaseelan[b]

[a]*Research Scholar, Department of Computer Science, Bishop Heber College Tiruchirappalli-620017, Tamilnadu, India.*
[b]*Head, Department of Computer Science, Bishop Heber College Tiruchirappalli-620017, Tamilnadu, India.*

Abstract

Data mining is an important research area in computer science. It is a computational process of determining patterns in large data. Image mining is one of important techniques in data mining, which involved in multiple disciplines. Image Classification Refers the tagging the images into a number of predefined sets. It's also includes image preprocessing, feature extraction, object detection, object classification, object segmentation, object classification and many more techniques. Image classification to produce the accurate prediction results in their target class for each case in the data. It is a very predominant and challenging task in various application domains, including video surveillance, biometry, biomedical imaging, industrial visual inspection, vehicle navigation, remote sensing and robot navigation. The aim of this study compares the some predominant data mining algorithms in image classification. For this review SVM, AdaBoost, CART, KNN, Artificial Neural Network, K-Means, Chaos Genetic Algorithm, EM Algorithm, C4.5 algorithms are taken.

Index Terms: Data Mining, Image Classification, Data Mining Algorithm, Kappa Coefficient, Classification Accuracy.

1. Introduction

Data mining is a dominant technology with prodigious potential to help the organization. The data mining tools predict the future trends, behaviors, knowledge driven decision. Data mining is a process of extracting the valuable information from large amounts of data. In other argument's data mining is mining the knowledge from data. The classification of data mining system is classified based on different criteria such as types of data and data models. Data mining makes classification models by using already classified data and finds the

* Corresponding author. Tel.: +91-9865381728
E-mail address: thamilselvan1987@gmail.com, jgrsathiaseelan@gmail.com

predicted pattern. The classification problems are used to identify the features of group in each case of class. Data mining can generate discover information and large number of rules. It should be applicable to any type of method, approaches and information repository. Data mining is used to be studying the relational databases, object oriented database, data warehouse, repositories, transactional databases, semi structured and unstructured. The data mining, classification methods generally used to classify the new objects. Image mining is one of the important research area in computer science it is a one part data mining. Image mining, classification is a machine learning classifier, it is trained by gave training samples. The trained classifier is able to classify the unlabeled or unknown vectors. The data mining methods are a core of KDD process such as classification, sequential patterns, and prediction models from different types of data.

Image mining is a calculation process of discovering and searching appreciated information in huge data. Image mining technique differs from standard mining [1]. Image mining technique is a very active research area for example, supervised image classification [2] and unsupervised image classification [3]. Image mining performs low level extracted features such as shape, color and texture. The extracted features denotes image content for image mining. In image classification, a classifier is trained based on giving training set. The trained classifier is used to classify unlabeled or unknown trained classes. Image mining is an extracting meaningful image content of huge image data set [4]. In this paper, we scrutinize the performance of a data mining algorithm based on the image classification technique.

Image classification process is automatically classify all the pixels in images which are used in hybrid manner and the method is based on nature of the data evaluated. Image mining handles extraction of knowledge and image data relationship in the images used to image retrieval, image processing, machine learning databases and data mining. Image mining techniques differ from image processing and low level computer vision because image mining is in extracting the patterns from a huge set of image data. The image processing methods and computer vision are in extracting particular feature from a single image. Image mining is nothing but applying existing data mining algorithms on images. There are some difference between image data base and relational database. The image mining techniques investigate the suitable framework for image mining. The image mining algorithms normally used in classification, clustering, image retrieval, image indexing, association rule mining, neural network and object recognition. Image classification is categorized into unsupervised classification and supervised classification. The supervised classification provides a collection of pre classified images.

Image mining appeals basic ethics from statistics, databases, pattern recognition, machine learning and soft computing. The data mining technique allows the use of data banks. Image mining is becoming an emerging research field in computer science because of the increasing large amount of data which indicates the new applications. For example, the use of high resolution satellite images permit the observation of slight objects. Image mining has two foremost subjects. The first one is mine the valuable information from large image data set and the second one combines a large collection of images associated with the image data. The image mining algorithms contain some following steps, i.e. feature extraction, object identification, record creation and association rule. This study represents the classification accuracy and the kappa coefficient accuracy of data mining algorithms using image data sets.

2. Related Work

Yang HongLei et al. [12] focused on remote sensing image classification by using EM (Expectation-Maximization) algorithm. The EM algorithm is used for the tasks in remote sensing image classification [5-11]. In remote image classification a unimodal assumption for conditional distribution is incongruous for high spatial resolution remote sensed images. The EM method is very sensitive for initializations in local minima. This method is mainly aimed to improve the classification accuracy of EM method. This proposed method shows the 83.8% accuracy in remote sensed image classification.

Mathanker et al. [13] concentrated on expanding the pecan defect classification accuracy by using the AdaBoost algorithm. The AdaBoost method performs well in selected classification accuracy. This work

indicates that the AdaBoost classifier is suitable for real time applications. This method is also extendable in new cultivation classification tasks. The Adaboost classifier method performs well in poor marking then accurately working for pecan defect classification. This method spectacle 92.2% accuracy in pecan defect image classification.

Bárbara et al. [14] focused on land cover image classification by using C4.5 method. This work is mainly aimed to increase the performance of image classification of urban land cover maps. The C4.5 method working in following ways: each and every node matches a value range of the attribute. The estimated attribute value is defined by root of each node. More data are needed to characterize the data. The C4.5 method automatically eliminates the avoidable nodes trough the pruning procedure. This technique provides 84% accuracy in image classification.

Helio et al. [15] deliberated on introduction about to Classification And Regression Tree (CART) and improve the performance of classification accuracy AVIRIS and Landsat digital images. The decision tree classification is a non-parametric method in pattern recognition. The CART includes identification and construction of decision tree by using training sample data for the correct classification purpose. CART method is a machine learning approach. The classification and regression tree can have a root node. The root node is divided into two sub node. The sub node can have grand sub node. This process stopped when no further split are not possible due to the lack of data. This method shows 92.9% classification accuracy in remote sensed digital image.

Bhuvaneswari et al. [16] described feature selection through genetic algorithm and the classification is done through KNN, NN, MLPNN, and J48 methods to classify the lung image dataset. This system has been tested with lung images and it achieved satisfactory results in lung diseases image classification. The genetic algorithm is based on population search method its travels from one set of point to another in a single iteration. The GA method involves three operations: crossover, mutation and selection. This method shows 90% accuracy in classification in lung image data sets.

Rajput et al. [17] focused on improving the classification performance by using K-Means method in color retinal images. The K-Means classification algorithm is superior case of general hard classification algorithms. With this technique is experimented on 100 images collected from various centers in Karnataka. All the images are visible optic disk. This work shows better accuracy in classification by using a color retinal image.

Begum et al. [20] concentrated on improving the support vector machine classification accuracy as well as to minimize the computational processing time of support vector machine. The proposed method is achieved by applying support vector machine training to the whole original training data. The support vector machine is a popular kernel based algorithm in image classification also, it can provide better accuracy in image classification [18-19]. This proposed technique is more gainful in lowest amount of training data. The low resolution images are classified using Support vector machine. This method increases the classification accuracy and also reduce the classification time. This method shows 97.15% classification accuracy in multispectral images.

Guo Yiqiang et al. [21] proposed a method Chaos Genetic Algorithm to increase the accuracy of remote sensing images. The CGA technique used for remote sensing image classification in the optimization problem. The optimization problem is solved by the GA method introduce chaos genetic algorithm. This method increases the optimal extend of population in the starting stage of the method. The remote sensing technology can increase the image classification accuracy. This method is divided into supervised methods and unsupervised method in image classification. The Chaos Genetic Algorithm technique achieved 90.1% accuracy in remote sensing image classification.

Ravi Babu et al. [22] provided an efficient and reliable method for handwritten digital image classification by using K-Nearest Neighbor algorithm. The proposed method tested using 5000 images to find classification rate. This proposed KNN classification method for classifying handwritten digit images using training database. The processing of k nearest neighbour method defines the computational classification time is done through learning technique. Normally k-nearest neighbour has two types of learning techniques, i.e. lazy based and instance based learning techniques. The K-NN method simple and easy to classify because computational

process is simple. This proposed method generates 96.94% accuracy in classification by using handwritten images.

Chuan-Yu Chang et al. [23] focused on improving the classification performance by using Radial Basis Function in texture images. The RBF neural network has many characters like approximation, robustness and simplicity. This is a successful method in image classification. In this method, they first fragmented, each texture image into sub images and then the sub images decomposes through wavelet transformation. The accuracy of the proposed method is compared with other techniques in the texture image classification. This RBF method shows better accuracy in image classification than other texture classification [24].

Soranamageswari et al. [25] presented experimental method for image spam classification by using Artificial Neural Network. It is an effective method in image classification for finding and solving feature extraction problems. For this exercise Back Propagation Neural Network (BPNN) is used. The BPNN is used to solving various problems. The Artificial Neural Network can have three types of layers basically such as input layer, hidden layer and output layer. The neurons are interconnected into the above mentioned layers. Each neuron is adjusting the weights automatically during the training process. The actual results are compared with goal value to find the performance of classification. This method shows 93% accuracy in image classification by using spam images.

Min Han et al. [26] found the good classification method which could achieve good classification accuracy to deal with remote sensing image classification. This author proposes a new classification method extreme learning machine (ELM). The ELM technique is feed forward neural network. To elucidate the training sample problem are accessible in remote sensing image classification. The extreme learning machine having three layers such as input layer, hidden layer and output layer. This proposed method approximately shows 90% accuracy in image classification by using remote sensing image classification.

Amini et al. [27] presented an image classification of hyperspectral images using Random Forest (RF) algorithm. The classification work is carried out based on unlabelled and labelled data. The RF classifier method has been proposed recently for image classification. This method is an ensemble classification technique that including collection of classifiers. This classifier uses a big number of separate decision trees for performing the classification. The overall accuracy achieved by random forest technique 73.58% and semi supervised random forest method achieved 82.63% accuracy in image classification.

Kersten et al. [28] focused on improving the classification accuracy by using Fuzzy C-Medians method. This method is calculated by sample values and associated with each value. The main advantage of this method reduces processing time and space complexity is parallel. This proposed system shows better accuracy in image classification by using POLSAR images.

Yan Wang et al. [29] presented data clustering algorithm for image classification in Landsat images. The data clustering method based on artificial neural network and fuzzy interference. It passes the training data in short period time. This method is applied in Landsat images to find the image classification accuracy. This method predicts 88.64% accuracy in image classification.

3. Comparative Analysis

In this section, tested datasets and comparative results for the data mining algorithms are to be discussed. This comparative study describes the purpose and limitations of data mining algorithms.

3.1. Data Set

To find image classification performance of data mining algorithms are tested in various image data sets. The used data sets are shown in table 2.

3.2. Comparison Between Data Mining Algorithms

This part describes the merits and demerits of algorithms are taken from this survey. Table 1 describes, the issues of data mining algorithm which is algorithms taken from this survey.

Table 1. Issues of Data Mining Algorithms

		Description of Algorithms	
S. No	Algorithms	Purpose	Limitations
1	SVM	• Most Effective methods in classification, especially popular in text classification • Compare with ANN, it captures the essential characters of the data. • Provides high accuracy in classification	• More complex to classify • Difficult to interpret for solving, parameter model • Several key parameters are needed to achieve the best classification result
2	AdaBoost	• AdaBoost powerful method of classification • The number of boosting rounds is available in the training process • Fast, Simple, and Flexible	• Very complex in noisy data and outliers • Not as robust in predicting the error • Fails to handle the weak learners if the rate is larger than 1/2
3	CART	• Handles missing values automatically • No probabilistic assumpt3ions • Variable selection performs automatically	• Step by step function, not a continuous one. • Poor modelling in a linear structure.
4	KNN	• Simple, effective, easy to implement and non-parametric • Provides low error rate in training process.	• Slow Process • Classification time is long • Difficult to find optimal value
5	Neural Network	• In complex domains it provides good result • Better for continuous domain • The testing process is fast	• Slow process in training
6	K-Means	• Low computational complexity • It handles large scale data sets • Simple and easy to implement	• Depends on several parameters. • This method doesn't have guarantee optimal solution • It fails in nonlinear data sets.
7	Chaos Genetic Algorithm	• Easy to solve optimization problem • It can solve non continuous, non-differential and multi-dimensional parameter problem. • Very fast and easy to implement • Easy to transfer in existing models and simulation	• Difficult to find a global optimal solution • This method cannot solve certain optimization problems.
8	EM Method	• Simplicity of representation • It shows sufficient high amount of un labelled data.	• Not a guarantee optimal solution
9	C4.5	• Suitable for real world problems. • Handles missing values • Split the data most accurately	• More simple rules. • High training samples are needed. • Unsatisfactory in practical application

3.3. Discussions

The performance data mining algorithm classification accuracy, used data sets and kappa coefficient are shown in table 2. From this study, each data mining algorithm essential in any one way. The classification accuracy results contain true positive values and true negative values. The classification performance calculated by the following formula.

$$Accuracy = \frac{\text{Number of true positive} + \text{Number of true negative}}{\text{Number of true positives} + \text{False positives} + \text{False negatives} + \text{True negatives}}$$

(1)

Table 2. Data Mining Algorithms Performance

			Performance of Algorithms		
S. No	Algorithms	Data Sets	Correctly Classified Instances in %	Incorrectly Classified Instances in %	Kappa Coefficient in %
1	SVM	Hyperspectral Images	97.15	2.85	94
2	AdaBoost	X-Ray Image of Pecan Defect	92.2	7.8	89
3	CART	Remote Sensed Digital Images	92.9	7.1	89
4	KNN	Handwritten Digital Images	96.94	3.06	93
5	Neural Network	Spam Images	93	7	-
6	K-Means	Color Retinal Images	Better Accuracy	-	-
7	Chaos Genetic Algorithm	Remote Sensing Images	90.1	9.9	87.63
8	EM Method	Remote Sensing Images	83.8	16.2	80.37
9	C4.5	World View-2 Images	84	16	82.34

This is describes study of data mining algorithms are used in different type images to find the image classification accuracy. The data mining algorithms are developed for various purposes. In this study, the data mining algorithms are taken for this review such as SVM, Adaboost, CART, KNN, Artificial Neural Network, K-Means, Genetic Algorithm, EM Method and C4.5 algorithm. From this study support vector machine and K-Nearest Neighbour shows better accuracy in image classification when compared with other data mining algorithms. The performance of data mining algorithms in image classification is shown in following Fig 1. The performance of a data mining algorithms in kappa coefficient is shown in fig.
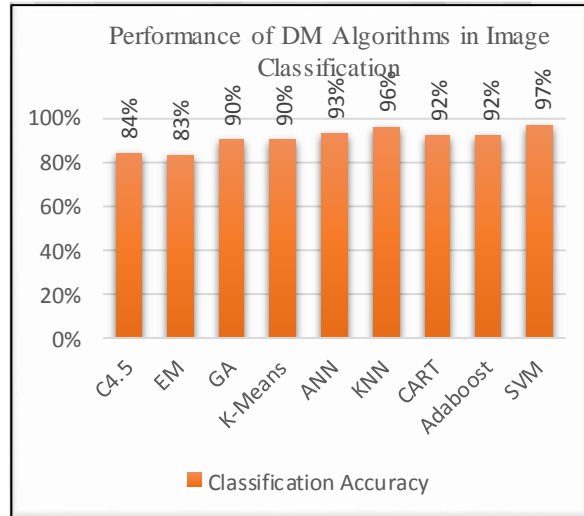
Fig.1. Image Classification Performance of Data Mining Algorithms

The fig 1 shows the classification accuracy performance by using different data mining algorithms. The algorithms taken for this review support vector machine shows the highest accuracy in image classification. The Fig 2 shows the performance of data mining algorithms by using kappa co efficient accuracy. From figure 1 support vector machine show highest accuracy in kappa co efficient accuracy.
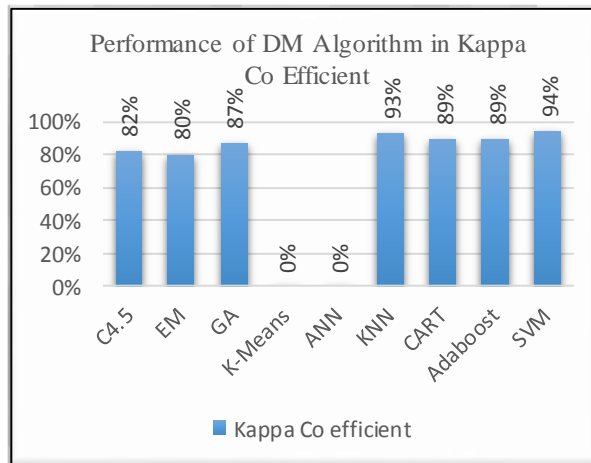

Fig.2. Kappa Coefficient Performance of Data Mining Algorithms

4. Conclusion and Future Work

In this review, we consider the performance of data mining algorithms in image classification which is analyzed based on classification accuracy and kappa coefficient. The support vector machine shows 97.15 % accuracy and k nearest neighbour shows 96.94 accuracy in image classification. From this review, Support Vector Machine and K-Nearest Neighbour show higher accuracy in image classification. Hence, we suggest

that SVM and KNN are the most predominant data mining algorithms in image classification. In future we planned to hybrid any two data mining algorithm to improve the classification accuracy and reduce the error rate.

References

[1]   William I. Grosky. "Managing multimedia information in database systems," Communications of the ACM, 40 (12) pp 72–80, 1997.
[2]   Davide Agnelli, Alessandro Bollini, Luca Lombardi. "Image classification: an evolutionary approach" Pattern Recognition Letters, 23, pp 303–309, 2002.
[3]   Yixin Chen, James Z. Wang. "Image categorization by learning and reasoning with regions" Journal of Machine Learning Research, 5, pp 913– 939, 2004.
[4]   Aura Conci, Everest Mathias M.M Castro. "Image mining by content" Expert Systems with Applications, 23, pp 377–383, 2002.
[5]   Bruzzone L, Prieto D.F. "Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images" IEEE T Geosci Remote S, 39 pp 456–460, 2001.
[6]   Figueiredo M.A.T, Jain A.K. "Unsupervised learning of finite mixture models" IEEE T Pattern Ana, 24 pp 381–396, 2002.
[7]   J.C Luo, Q.M Wang, J.H Ma, Y Liang, C.H Zhou. "The EM-based maximum likelihood classifier for remotely sensed Data" Acta Geod E 31, pp 234–239, 2002.
[8]   Chakravarty S, Qian Du, Hsuan Ren. "Adaptive Gaussian mixture estimation and its application to unsupervised classification of remotely sensed Images" Geoscience and Remote Sensing Symposium, IGARSS'03. Proceedings. France: IEEE International 3, pp 1796–1798, 2003.
[9]   Kersten P.R, Jong-Sen Lee, Ainsworth T.L. "Unsupervised classification of polarimetric synthetic aperture radar images using fuzzy clustering and EM clustering" IEEE T Geosci Remote S 43, pp 519–527, 2005.
[10]   Thales Sehn Korting, Luciano Vieira Dutra, Guaraci Jose Erthal, Leila Maria Garcia Fonseca. "Assessment of a modified version of the EM algorithm for remote sensing data classification" Lect Notes Comput Science 6419, pp 476–483, 2010.
[11]   Yang H L, Peng J H, Li S H, et al. "Log-principal component transformation based EM algorithm for remote sensing classification" Acta Geod E 39, 378–382, 2010.
[12]   Yang HongLei, Peng JunHuan, Xia BaiRu, Zhang DingXuan. "An improved EM algorithm for remote sensing classification" Springer, Chinese science bulletin Vol.58 No.9 doi: 10.1007/s11434-012-5485-4, pp 1060-1071, 2012.
[13]   S.K. Mathanker, P.R. Weckler, T.J Bowser, N. Wang, N.O. Maness. "AdaBoost classifiers for pecan defect classification" Elsevier Computers and Electronics in Agriculture 77, pp 60–68, 2011.
[14]   Bárbara Maria Giaccom Ribeiro, Leila Maria Garcia Fonseca. "Urban Land Cover Classification using WorldView-2 Images and C4.5 Algorithm" IEEE proceeding of the JURSE, pp 21-23, 2013.
[15]   Helio Radke Bittencourt, Robin Thomas Clarke. "Use of Classification and Regression Trees (CART) to Classify Remotely-Sensed Digital Images" Geoscience and Remote Sensing Symposium, IGARSS '03. Proceedings. IEEE International (Volume: 6) pp 3751 – 3753, 2003.
[16]   C. Bhuvaneswari, P. Aruna, D. Loganathan. "A new fusion model for classification of the lung diseases using genetic algorithm" Elsevier, Egyptian Informatics Journal, Volume 15, Issue 2, pp 69-77, 2014.
[17]   Dr. G. G. Rajput, Preethi N. Patil. "Detection and classification of exudates using k-means clustering in color retinal images" Fifth IEEE International Conference on Signals and Image Processing DOI 10.1109/ICSIP.2014.25 pp 126-130, 2014.
[18]   F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines" IEEE Transaction On Geoscience and Remote Sensing, vol. 42, no. 8, pp. 1778-1790, 2004.

[19] G. Camps-Valls, and L. Bruzzone, "Kernel-based methods for hyperspectral image classification" IEEE Trans. on Geoscience and Remote Sensing, vol. 43, no. 6, pp. 1351-1362, 2005.

[20] Begum Demir, S. Erturk. "Improving svm classification accuracy using a hierarchical approach for hyperspectral images" 16th IEEE International Conference on Image Processing ISSN: 1522-4880 pp 2849-2852, 2009.

[21] Guo Yiqiang, Wu Yanbin, Ju Zhengshan, Wang Jun, Zhao Luyan. "Remote sensing image classification by the Chaos Genetic Algorithm in monitoring land use changes" Elsevier Mathematical and Computer Modelling 51, pp 1408-1416, 2010.

[22] U Ravi Babu, Y. Venkaswarlu, Aneel Kumar Chintha. "Handwritten Digit Recognition Using K-Nearest Neighbour Classifier" IEEE World Congress on Computing and Communication Technologies ISBN: 978-1-4799-2876-7 pp 60-65, 2014.

[23] Chien-Cheng Lee, Pau-Choo Chung, Jea-Rong Tsai, Chein-I Chang. "Robust Radial Basis Function Neural Networks" IEEE Trans. on Neural Networks, vol. 29, no. 6, 1999.

[24] Chuan-Yu Chang, Shih-Yu Fu. "Image Classification using a Module RBF Neural Network" IEEE Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06) ISBN 0-7695-2616-0, pp 270-273, 2006.

[25] M. Soranamageswari, C. Meena. "Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Networks" Second IEEE International Conference on Machine Learning and Computing, ISBN: 978-1-4244-6007-6, pp 101-105, 2010.

[26] Min Han, Ben Liu. "Ensemble of extreme learning machine for remote sensing image classification" Elsevier Neurocomputing, Volume 149, Part A, 3 pp 65-70, 2015.

[27] S. Amini, S. Homayouni, A. Safari. "Semi-Supervised Classification Of Hyperspectral Image Using Random Forest Algorithm" IEEE International on Geoscience and Remote Sensing Symposium (IGARSS), pp 2866-2869, 2014.

[28] P. R. Kersten, J. S. Lee, T. L. Ainsworth. "Classification of POLSAR Images using a Fast Fuzzy C-Medians Clustering Algorithm" IEEE International on Geoscience and Remote Sensing Symposium, IGARSS '04. Proceedings. Volume: 1 ISBN: 0-7803-8742-2 pp 552-555, 2004.

[29] Yan Wang, M. Jamshidi, P. Neville, C. Bales. "Multispectral Landsat Image Classification Using A Data Clustering Algorithm" International IEEE Conference on Machine Learning and Cybernetics, Volume: 7, ISBN: 0-7803-8403-2 pp 4380-4384, 2004.

Author(s) Profiles

Mr. P. Thamilselvan, received his M. Phil. Degree in Computer Science from Bharathiddasan University, Tiruchirappalli, Tamilnadu, India in 2013. He also received M.C.A. Degree in Computer Applications from Anna University, Chennai, Tamilnadu, India in 2012. He is currently pursuing the Ph.D. in Computer Science from Bharathidasan University, Tiruchirappalli Taminadu, India. His area of research interests includes Artificial Neural Networks and Image Mining. E-mail: thamilselvan1987@gmail.com

Dr. J. G. R. Sathiaseelan is the Head of Computer Science Department at Bishop Heber College, Tiruchirappalli. He has 25 years of teaching experience. He has presented more than 20 research papers in International conference publication which are published in IEEE, ACM, Springer, and reputed journals. Dr. Sathiaseelan has authored a book entitled as, "Programming In C#, .Net", which was published in PHI, New Delhi, in 2009. His research areas include Web Services security, Data mining, image processing and big data analytics.
E-mail: jgrsathiaseelan@gmail.com