iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IX **Month of publication:** September 2024

DOI: https://doi.org/10.22214/ijraset.2024.64224

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IX Sep 2024- Available at www.ijraset.com

Supervised, Unsupervised Learning Methods in Data Mining: A Comprehensive Survey

S.Sivasankari[1], DR.S.Sukumaran[2]
[1]*Ph.D Research Scholar,* [2]*Associate Professor*
[1,2]*Erode Arts and Science College, Erode, Tamilnadu, India*

Abstract: ***Data mining is the area of computational science consider as the subset of Artificial Intelligence, Machine Learning that concentrate on cognizing patterns and structures in data to promote learning, analyzing, and decision making beyond human interaction. In simplest term, data mining allows the user to feed data mining algorithms on an immense amount of data and then make the computer to make data-driven recommendations and decisions based on the input or training data. Data is the essential component of all business. Data mining with supervised and unsupervised methods is the key to unlock the value of business data and keep ahead of the competition. This paper presents a comprehensive survey on Data mining process, supervised, unsupervised methods, through a survey of literature considering the significant research areas using classification, clustering, and association rule mining.***
Keywords: Machine learning, Data mining, Supervised learning, Unsupervised learning.

I. INTRODUCTION

Data mining is the procedure to find patterns, [7] anomalies, and correlations in a large data sets in order to predict outcomes. With the use of broad range of techniques, the process of digging knowledge from data is done to discover hidden connections and predict future trends. Hence, it is referred as "Knowledge Discovery in Databases (KDD)". The advances in processing power and speed transform the nature of analyzing to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. [36] The increasing complexity of the data leads to higher potential to uncover relevant insights. Data mining is at the heart of analysis across a variety of industries and disciplines. . The field comprises of three scientific disciplines namely Statistics, Artificial Intelligence, and Machine Learning.

1) *Classical Statistics*- It forms the basis of most technology on which data mining is built. The techniques include regression analysis, standard deviation, standard variance, standard distribution, cluster analysis, discriminatory analysis, and confidence intervals.
2) *Artificial Intelligence* - This is based on heuristics, which is opposed to statistics. It applies human- based thoughts to statistical problems. Some high-end commercial products, such as query optimization modules for Relational Database Management System, adopted a specific AI concept.
3) *Machine Learning*- Machine learning is combined with statistics and AI. [35] It is an evolution of AI since it mixes AI heuristics with statistical analysis. Machine learning tries to enable computer programs to know about the data that make a decision based on the characteristics of the data being examined. Hence, it uses statistics for basic concepts and adds more AI heuristics to accomplish its target. It is categorized into two types viz. supervised and unsupervised in which the former relies on trained data, the later relies on unlabeled set.

Modern-day companies cannot live without data. [13] To stay ahead of the competition with smart decisions, they have to evolve and keep up with technological evolution and upcoming digital trends. Therefore, businesses today are prioritizing staying ahead of all the new developments in the field of data science and analytics with data mining techniques. Some most important data mining applications include Financial Analysis, Telecommunication Industry, Intrusion Detection, Retail Industry, Higher Education, Energy Industry, Spatial Data Mining, Biological Data Analysis, Scientific Applications, Manufacturing Engineering, Criminal Investigation, Counter terrorism etc.

A. Data Mining Process, Methods

Data mining (KDD) process comprise of Data selection, Preprocessing, Transformation, Mining, Evaluation and Knowledge gaining while the CRISP-DM (Cross-Industry Standard Process for Data Mining) process comprise of Business understanding, data understanding, Data preparation, Modeling, evaluation and Deployment.

The mining methods are classified into three types [15] namely supervised with labeled data, unsupervised without label, and semi-supervised that has partially labeled with more unlabeled data.

1) *Supervised learning*- It has labeled datasets to train algorithms to [33] classify data or predict outcomes accurately. The dataset is split into training, testing also validation. The training dataset includes inputs and correct outputs. The input data is fed into the model, it adjusts its weights till the error has been reduced properly. Supervised learning helps organizations solve for a variety of real-world problems at scale. It can be separated into two types of problems namely classification and regression. The former uses an algorithm to accurately assign test data into specific categories where as the later are used to understand the relationship between dependent and independent variables.
   Algorithms: Neural networks, Decision tree based learning, Naive bayes, K-nearest neighbor, Logistic regression, Linear regression.
2) *Unsupervised learning* - Unsupervised learning methods uses unlabeled [9] data from which it discovers patterns. The goal is to find the underlying structure of data, grouping data according to its similarities, and represent that in a compressed format. Clustering, Association rule mining belongs to the category of unsupervised algorithms.
   Clustering Algorithms : Hierarchical Clustering with Agglomerative, Divisive. Partitioning Clustering with K-Means, K-Medoids, K-Mode, K-Median, X-Means, Fuzzy C-Means. Probability based clustering namely Expectation Maximization. Density based clustering DBSCAN.
   Association rule mining algorithms: Apriori, FP-Growth.
3) *Semi-supervised learning* - This lies between the Supervised and Unsupervised learning methods. [23] Training data is a combination of both labeled and unlabeled data. However, labeled data exists with a very small amount while it consists of a huge amount of unlabeled data. To work with the unlabeled dataset, there must be a relationship between the objects with assumptions.
   Assumptions: Continuity assumption, Cluster assumption and Manifold assumption.

II. RELATED WORKS WITH VARIANT DATA MINING METHODS

The recent research papers that implemented data mining classification, clustering, association rule mining methods for variant applications with its process, limitations is discussed in this section elaborately.

A. *Supervised Learning methods – Classification, Regression*

Anjuman Prabhat *et al* [2] compared two machine learning classification algorithms naive bayes and logistic regression for sentiment classification on movie reviews. TF-IDF method with term/word level sentiment classification was carried out in MapReduce and Hadoop Distributed File system for categorizing twitter movie reviews with two polarities positive and negative. The resultant values of these two algorithms showed that logistic regression has better accuracy level with 76.76 % and naive bayes has 66.66 % for 6 MB dataset. However, the accuracy is inadequate.

Sanditya Hardaya *et al* [30] designed a classification model Support Vector Machine (SVM) for the classification of the complaints and proposals using text mining technique. In preprocessing, stemming, case folding, filtering, synonym recognition and the word vector was created using Term Fequency (TF), Inverse document frequency (IDF) and Singular Value decomposition (SVD). The performance was evaluated and found that classification model with stemming, synonym recognition was the most accurate among others with 91.37% accuracy rate. Even though the method TF-IDF can be tried to improve the accuracy.

Ayesha Anzer *et al* [5] implemented a linear regression approach to predict final exam scores from early course assessments of the students. The datasheet comprise of different attributes like gender, quiz 1, quiz 2, quiz 3, quiz 4, assignment 1, assignment 2, assignment 3, midterm, and final exam grades. In preprocessing, unneeded attributes are removed and missing values were modified then linear regression model was applied. Finally, it was noted quizzes and total assignment score was most important factor to predict the academic performance from the values t-statistics with 3.34 for Quiz 2 and p-value with 0.64 for Quiz 1.

Chittem Leela Krishna *et al* [6] applied Deep Neural Net MLP (feed forward) classifier for patient analysis of Cleveland Heart Disease data from UCI. Initially, data preprocessing is done with inter quartile range, missing value is replaced, best first search method was used to have subset of features and the five class is transformed to 2 (binary) class. The proposed deep neural network consists of three hidden layers with the size 15, 12, 5, sigmoid activation function, and mean absolute error as a loss function. The Learning and momentum rate was 0.15 and 0.2.

From the results, it was found that the proposed model deep neural network obtained 85.4 % accuracy and it is followed by support vector machine 84.4 %, naive bayes with 83.8 %, decision tree with 80.8 %.

Najat Ali *et al* [19] put forth a combination of similarity measure in K-nearest neighbor to investigate the performance on heterogeneous datasets The different domains are Hypothyroid, Hepatitis, Treatment, Labour training evaluation, Catsup, Azpro.
For the analysis, Euclidean, Manhattan, combined measures namely Euclidean-Jaccard, Cosine-Jaccard, Canberra-Jaccard was used in R. In preprocessing, irrelevant features are removed, missing values were replaced and numerical features were normalized. The training and testing set consists of 80%, 20 %. The overall results proved that the combination of discussed measures gave a promising result with the *K* value 1. Among all the datasets, the discussed combined measures gave 100 percent accuracy for hypothyroid dataset. Still, the work does not prove the accuracy with minimized processing time.
Farhan Ullah *et al* [10] proposed a Deep feed forward Neural Collaborative Filtering for educational services recommendation. The experiments was performed in keras on the benchmark good books 10k. Initially, the users' identifier and books identifier features are mapped into N-dimensional vectors and then this array is feed into multilayer perceptron. Variant activation functions tanh, Elu, sigmoid was compared with the proposed function Relu. Four layer-hidden architecture is employed with the neurons (32, 64, 128, and 256). From the results it was observed that the model with 4 hidden layers, L2 regularization, Huber loss function, Relu activation function, Adam optimizer gave less RMSE-0.82) and Mean absolute error (MAE-0.60) and outperforms the existing method. Nevertheless, processing time for each hidden layer structure with the activation function must be noted.
Lalit Mohan *et al* [17] implemented support vector machine (SVM) with variant kernels radial basis kernel, linear kernel, polynomial kernel, sigmoid kernel on iris dataset to assess the performance. From the initial results it was observed, SVM with RBF kernel gave high accuracy with 97.3 % and it followed by linear kernel (96.66%), polynomial (95.33%) and sigmoid (88.66%). Then, the parameters in the kernel selection cost, epsilon was modified in the RBF kernel to have further improvement in the accuracy and the accuracy was reached from 97.3 % to 98.6%. However, the proposed method takes longer time.
Niko Reunanen *et al* [20] proposed two ensemble based outlier detection method KNN with LOF and logistic regression with LOF in heterogeneous datasets to handle bias and variance in outlier analysis. The work applied data bagging and feature bagging. In first strategy KNN and LOF were chosen for outlier detection with optimized outlier ensemble (OOE). KNN was applied with the distance functions Manhattan, Euclidean, Chebyshev, Cosine, Correlation, and Canberra. In second strategy, LOF with logistic regression was chosen. The experiment was done with benchmark datasets and found that OOE method found 50 % outliers with best AUC 0.816. Still only one score AUC is used hence other parameters can be used.
Raed Alsini *et al* [27] put forth a method Isolation Forest based on a Sliding window (SW) with LOF (IFS-LOF) to quantify the anomalies and outliers in concrete mixtures. Different sizes of windows 100, 200, 300, 400 were used to evaluate the performance of LOF alone, LOF with sliding window LOF-SW and LOF with Isolation forest IFS-LOF. From the listed value in the results, it was noted that for 100, 200 size window LOF gave high accuracy 90.89%, 96.68% respectively, for 300 LOF-SW gave high accuracy 95.38%and 400 IFS-LOF gave high accuracy 97.48%. Hence, the proposed method IFS-LOF needs large window size. However, the method IFS-LOF takes high processing time.
Wenxian Feng *et al* [38] designed an improved random forest algorithm FSRF for six types of datasets in UCI. Feature selection methods filter method with embedded method and wrapper method was used. This subset was applied to random forest algorithm. The training and testing set is split into 70:30 and six types of datasets from UCI (LSVT, srbct, Arcene, gistte_train, Parkinson, cleveland, ionosphere) were taken for analysis. From the observed values, it was noted that filter- wrapper combined method with random forest gave high accuracy for LSVT 89%, srbct 100%, arcene 97% and ionosphere 95%. The overall results showed the combined feature selection method proved to be the best but, the time complexity of the algorithm is increased.
Archana R. Panhalkar *et al* [3] introduced Tangent Weighted Decision Tree Forest (TWDForest) C4.5, for heterogeneous dataset. It creates a weighted decision tree forest. Along with that ensemble method, bootstrap sampling was used. The weight of 1.0 was assigned to each attribute. The method was assessed with 15 datasets taken from UCI repository. The work was also evaluated for variant number of trees and noted the proposed ensemble TWDForest with 100 trees produced better accuracy 98.4%. But, the processing time is very long as compared with random forest.
He Xu *et al* [11] introduced a novel outlier detection algorithm KNN-LOF for activity recognition system based on a multisensor. Initially, k-nearest neighbors algorithm is applied to divide different areas for outlier attributes then, a hierarchical adjacency order is proposed. Hence, the reachable distance of an object was redefined with the average sequence distance in the hierarchy. The work was assessed with different methods LOF –Local outlier factor, RDOS- relative desity based outlier, and COF-connectivity based outlier factor with *K* the values 20, 45. For the value 45, the proposed method KNN-LOF found 101 outliers and has high area under curve value 0.96. But still, the time complexity is high.
Xiaoyu Luo *et al* [39] implemented Support Vector Machines (SVM) model in classifying English text, documents. The samples were collected from the UCI repository.

The work was experimented in WEKA tool for four type of datasets with SVM, NB, and LR. From the listed values. the proposed method SVM gave high precision values 0.88, 0.76, 0.63 for the datasets 1, 2, 3 respectively. But still, the precision can be improved also processing time for each classification method might be noted.

B. Unsupervised Learning methods – Clustering

Pedro albuquerque *et al* [24] proposed a customer segmentation model for real Mobile TV service through Support Vector Clustering (SVC). SVC based on support vector machine has a kernel trick and this work used inner product kernel. The data were initially standardized to possess zero mean and unit variance. The proposed method was assessed against hierarchical agglomerative with the linkage criteria ward, single, complete, average, mcquitty, meadian, centroid and found that the proposed method SVC has less sum of squares i.e 1.06. Still, the processing time is not listed for all the discussed methods.

Raja Kishor *et al* [28] presented a hybridization technique of Expectation-Maximization (EM) and K-means (KM) methods on benchmark datasets. These two algorithms was made hybrid in the order EM, KM. For the assessment, three datasets from UCI namely Magic Gamma, Poker Hand, Letter Recognition, three synthetic datasets were used. Different *K* values starts from 10 -15 was applied. The minimum processing time 0.111 seconds was taken for magic gamma dataset with *K* =11. The minimum sum of squares was obtained for the *K* value 15 as 32.75 for Poker handset dataset. Though, variant distance measures or kernel function with K-means can be substituted for further improvement.

Saptarshi Sengupta *et al* [31] coupled Quantum-behaved Particle Swarm Optimization with the Fuzzy C-Means Clustering algorithm and was tested in Matlab on a suite of datasets such as iris, breast cancer wiscosin, seed dataset, mammographic mass, sonar dataset from the UCI Machine Learning Repository. The proposed method was compared with particle swarm optimization with K-means and Quantum particle swarm optimization with K-means and from the observed results it was found the proposed method has highest F-measure 0.9641 for breast cancer dataset. However, the processing time for each method is not noted.

Ping Yang *et al* [25] presented an improved method for outlier detection on benchmark datasets to handle time overhead in LOF. This paper applied a Neighbor Entropy Local Outlier Factor (NELOF). Firstly, Self-Organizing Feature Map (SOFM) algorithm was improved to ease the calculation of each data point's local outlier factor . Then, K-distance neighborhood was replaced with relative K-distance neighborhood to redefine the local outlier factor. SOFM clustering was improved with canopy initialization for neurons. The work was conducted using python on seven different datasets syn_data1 and syn_data2, Iris, Wine, Glass, TEBHR, LDPA from UCI. The improved SOFM gave high cluster quality 0.97 and then the method NELOF gave high accuracy outlier detection 98.7 % with less execution time around 0.8000 sec. Still, the work can be experimented for high dimensional data.

Praveen *et al* [26] analyzed three linkage criteria on benchmark dataset using Hierarchical agglomerative clustering such as single, complete, average, and average weighted. Dataset was taken from UCI repository and evaluated using JAVA programming language. From the results, it was observed that single linkage takes least time i.e. 3591 ms. However, the details of dendrogram and number of clusters formed is not focused.

Rishabh Ahuja *et al* [29] built a movie recommender system with K-Means Clustering and K-Nearest Neighbor algorithm. The movielens dataset was taken from kaggle and implemented in python programming language. In preprocessing step, utility matrix was created. Then K-means algorithm by considering within Sum of Squares is implemented for right number of clusters. Then, for prediction for top *n* movie rating, KNN algorithm was applied with the Pearson coefficient similarity. In final it was noted even for large number of clusters (68) the RMSE value is 1.231. But, in existing the same value was obtained for 19 clusters. However, within Sum of Squares is applied for choosing number of clusters but its value is not noted.

Iulia-Maria Radulescu *et al* [14] put forth a method to cluster short text documents using a modern document embedding model, specifically Doc2Vec with Distributed Memory Model (PV-DM), and Distributed Bag of Words Model (PV-DBOW) for DBSCAN and HDBSCAN on a set of 12263 articles extracted from the Arxiv database. Stemming, Lemmatization were used for preprocessing. The results showed that PV-DM embedding wordnet lemmatizer obtained 0.56 Adjusted RAND index for DBSCAN and 0.59 Adjusted Mutual Information values thus produce qualified cluster. However, number of clusters was not discussed.

Noor S. Sagheer *et al* [21] used the canopy clustering algorithm with k-means clustering algorithm for health insurance big dataset. Two types of mode was chosen serial in which K-means algorithm is implemented in a single machine and Parallel mode in which the data parallel divides the whole data into small subsets. Results were generated and found that the parallel mode for K-means with canopy took lesser time 1.048 hour and K-means without canopy took 1.847 hour thus proved the efficiency. But, the work concentrates on speed only, the quality of the cluster is not proved.

Shan-shan Li *et al* [32] put forth an improved DBSCAN based on neighbor similarity that utilizes Cover Tree. Datasets includes PAM4D, HOUSE, REACTION, MOCAP, and FONT from UCI repository for analysis.

This proposed method was compared with two earlier proposed one namely DBSCAN and ρ- approximate DBSCAN. Initially all missing values are set to zero, and all data value are normalized. The analysis was carried out for variant ε, min points.

From the results, it was proved that proposed method DBSCAN with cover tree speeds up the process by having least time 19.68 seconds for the REACTION dataset with Dimension 28, ε -5000 and minpts -50. Nevertheless, number of clusters and the quality of the cluster is not specified.

Joelson Antonio dos Santos *et al* [16] put forth a method that combine the quality of DBSCAN and hierarchical clustering on benchmark. It combined the aspects of density based clustering and hierarchical clustering that produce a complete density-based hierarchy clustering using Map reduce framework. Recursive Sampling Approach was created to draw sample from a whole dataset and then it was combined with a data summarization technique called data bubbles with parallelism. The proposed method proved with highest Adjusted Rand Index as 0.881 for all the datasets. Also, regarding runtime the hierarchical category agglomerative took minimum 2.11 min. which was less than the divisive. Memory utilization is not proved.

C. Unsupervised Learning methods - Association rule mining

Ashish Shah *et al* [4] proposed a modification to the apriori algorithm by using a hash function which divides the frequent item sets into buckets. Further, a novel technique was used in conjunction to eliminate infrequent itemsets from candidate set. This top down approach saves time and space. The proposed technique was explained with example contains 15 transactions, minsupport value 20%. The work was proved with number of scans that was reduced to one to get frequent itemset but it should be evaluated until the rules generated. Also, the processing time must be noted for each scan.

Andi *et al* [1] implemented FP growth for book search to get the association between each borrowed by the student. The results was analyzed for 1- frequent itemset, 2- frequent tiemset and 3-frequent itemset with support count and filter the largest count itemsets separately. With the support count 4, for 1-3 frequent itemset list, 21 rules were formed. But still, the algorism has to improved to get more number of rules.

Le Hoang Son *et al* [18] designed a new mining algorithm Association rule mining Apriori based on Animal Migration Optimization (AMO), called ARM-AMO, to reduce the number of association rules. For the analysis a supermarket data was taken and assessed with variant support and confidence in MATLAB tool. The results indicated that ARM-AMO provides better performance for the support, confidence values starts from 0.1 - 1.0. ARM –AMO generates rules around 3120-514 respectively, memory consumption around 41.2 MB - 3.1MB respectively and processing time around 312 sec. - 31 sec for the support and confidence values 0.1-1.0. Still, the work can be experimented with other parameters such as lift, leverage etc.

Dawei Dong *et al* [8] proposed a Improved Ant lion optimizer (IALO) with FP-Growth (FP) in nursery school dataset. The main idea of IALOFP includes Random walks of ants, Trapping Antlion's hole, Building trap, Sliding Ants towards Antlion, Catching Prey & Re-building the hole and finally Elitism with a good fitness that contains support and confidence. The results showed IALOFP produced 170 rules that was more than genetic algorithm, ALO alone methods. Nevertheless, the processing time for each method is not discussed.

Xu Hongfei *et al* [40] applied Apriori algorithm with undirected graph that speeds itemset generation instead of matrix to store itemsets and so candidate set is not needed. The proposed method used depth first search algorithm to traverse undirected itemset graph only once to acquire new frequent itemsets. The method was implemented on VC++ language with a sample data, and found that improved one need less processing time starts from 10 sec. to 25 sec. for the support count 0.3 to 0.1. However, the number of association rule is not focused.

Huan-Bin WANG *et al* [12] improved the apriori algorithm based on Map Reduce model with the idea of parallelism. First, local frequent itemsets on each sub node in the cluster are calculated and then merged into the global candidate itemsets. Finally, the frequent itemsets that meet the conditions were filtered based on minimum support threshold. Thus improved algorithm needs to scan the transaction database only twice. The execution process is verified by a specific example with 10 transactions but without the evaluation parameters such as number itemset generated, number of rules formed, processing time.

Norulhidayah Isa *et al* [22] applied FP-Growth to associate the relationship between the supply needed and the projects acquired for stopping maverick buying in electric company. The procedure implemented in Neddy Enterprise Sdn. Bhd (NE) electrical company. Initially, based on the purchase order, date, supply, quantity the data was consolidated and then FP-Growt with support 0.2, confidence 0.8 value the frequent itemsets and rules were formed in Rapidminer. However, number of rules for the prescribed values, processing time was not listed with consolidated dataset.

Table I. summarizes the recent works on supervised learning classification methods Naive bayes, Logistic regression, SVM, Linear regression, Deep learning multi feed forward neural network, KNN, SVM, Isolation forest, Random Forest, Decision tree C4.5 on the applications namely sentiment analysis, Text mining, education, Biological analysis, Outlier analysis, Text classification and some are implemented on heterogeneous data.

TABLE I. Summary of Related works on Supervised Learning method Classification

Author, Year	Method	Application	Merits	Demerits
Anjuman Prabhat *et al* 2017 [2]	Word vector: TF-IDF Classifier: Naive bayes, Logistic regression	Sentiment analysis	Categorize twitter movie reviews with two polarities positive, negative. Logistic regression got high accuracy with 76.76 %	Analysis based on processing time is not carried out.
Sanditya Hardaya *et al* 2017 [30]	Preprocessing: Stemming, Case folding, synonym recognition. Text mining Method: TF, IDF, SVD Classifier: SVM model	Text Mining	Precision reached to higher level 91.37% on community complaints and proposals dataset.	Spatial-Temporal details should be included to have in-depth analysis. method TF-IDF has more advantage but it is not used.
Ayesha Anzer *et al* 2018 [5]	Preprocessing: Removing irrelevant attributes, Replacing missing values. Classifier: Linear regression	Education	Statistical analysis was done to get the most important attribute with p-test value, T-test value using linear model.	Classification is not done to assess the students' performance.
ChittemLeela Krishna *et al* 2019 [6]	Preprocessing: Missing value replacement , Inter quarile range, Feature subset with best first search. Classifier: Deep multi feed forward network.	Biological analysis	The accuracy was 85.4% than other methods for Cleveland Heart Disease data.	Still, have to improve the accuracy by adjusting the parameters in the deep neural classifier.
Najat Ali *et al* 2019 [19]	Training - 80:20 Preprocess: Removal of Irrelevant features, replacing Missing values , Normalization. Classifier-KNN with combined similarity measure	Heterogeneous applications	Accuracy is high 100% for Hypothyroid dataset.	Accuracy is less for other applications.
Farhan Ullah *et al* 2020 [10]	Collaborative Filter. Classifier: Deep multi feed forward Neural network	Education Recommendation system	4 hidden layers, L2 regularization, Huber loss function, Relu activation function, Adam optimizer was utilized and obtained RMSE 0.82 on book dataset.	RMSE value ≥ 0.5 reflects the poor ability of the model.
Lalit Mohan *et al* 2020 [17]	Cross Validation Classifier: SVM with	Biological analysis	High accuracy 98.6 % was obtained on iris dataset.	Takes high processing time.

	RBF Kernel			
Niko Reunanen *et al* 2020 [20]	Outlier Factor: LOF Classifier: KNN with optimized ensemble	Outlier analysis	Handle bias and variance. AUC 0.816 on benchmark datasets	Finds 50 % outliers only.
Raed Alsini *et al[* 2020 [27]	Outlier Factor: LOF Classifier : Isolation Forest with sliding window	Outlier analysis	Outliers in concrete mixture dataset is found with varying window size and highest accuracy was 97.48 % for 400 size window.	Takes high processing time.
Wenxian Feng *et* al 2020 [38]	Training:70:30 Preprocessing: Filter-wrapper combined. Classifier: Random Forest	Heterogeneous applications	Combined feature selection method with the classifier proved to be the best with 100 % accuracy on benchmark dataset.	Time complexity of the algorithm is increased.
Archana R.Panhalkar *et al* 2021[3]	Classifier: Tangent Weighted Decision Tree Forest C4.5	Heterogeneous applications	Produced better accuracy 98.4 % on benchmark dataset	The processing time is very long.
He Xu *et al* 2021[11]	Outlier factor : LOF Classifier : KNN	Outlier analysis	High number of outlier 101 for the *K* value 45 on activity recognition system.	Neighbor search algorithms is not concentrated.
Xiaoyu Luo *et al* 2021[39]	Classifier: SVM	Text Classification	High precision value 0.88 was obtained for English text document.	Text preprocessing methods is not implemented.

Table II, summarizes the unsupervised learning clustering methods such as Support vector clustering, Expectation Maximization, K-Means, Hierarchical agglomerative, Fuzzy C-Means, Self Organizing Map, Density based DBSCAN and Canopy clustering on the applications namely customer segmentation, movie recommendation, text clustering, Health insurance and some are applied on heterogeneous benchmark datasets with its merits and demerits.

TABLE II. Summary of Related works on Unsupervised Learning method Clustering

Author, Year	Algorithm	Application	Merits	Demerits
Pedro albuquerque *et al* 2015 [24]	Preprocessing: Standardize. Clustering: Kernel based Support vector clustering with inner product kernel	Customer segmentation	Cluster distribution with 5 clusters is proved with less sum of squares 1.06 on Mobile TV service data.	Variant size cluster is not tried to get further least sum of squares.
Raja Kishor *et al* 2016 [28]	Partition K-Means with Probability based EM clustering.	Heterogeneous applications	Variant *K* value was tried.. Minimum processing time 0.11 sec. and least sum of squares 32.75 was obtained for *K* value 15.	K-Means with Euclidean cannot handle outliers.
Saptarshi Sengupta *et al* 2018 [31]	Quantum-behaved Particle Swarm Optimization with Partition based Fuzzy C-Means clustering	Heterogeneous applications	FCM helps to partition data based on membership probabilities. Breast cancer application has highest F-measure 0.9641.	Randomly chosen centroids leads to more iteration.
Ping Yang *et al* 2019 [25	Neural network based SOM clustering with LOF	Heterogeneous applications	SOM, Relative K-distance redefines the LOF. Cluster quality 0.97, outlier detection 98.7 %, less time 0.8000 sec on benchmark datasets was obtained.	The work is not proved for high dimension datasets.
Praveen *et al* 2019 [26]	Linkage based Hierarchical Agglomerative clustering	Heterogeneous applications	Analyzed all the linkage criteria single, complete, average, average	Details of dendrogram and

			weighted on benchmark dataset. Single linkage takes least time 3591 ms	number of clusters formed is not focused.
Rishabh Ahuja *et al* 2019 [29]	Partition clustering K-Means with supervised classification KNN	Movie recommendation system	Utility matrix with user and movie rating was created. Top rated Movies was listed. Large number of clusters (68) with RMSE value 1.231 was obtained on movie review dataset	Processing time for producing the large number of clusters is not noted.
Iulia-Maria Radulescu *et al* 2020 [14]	Density based clustering DBSCAN	Text clustering	Cluster short text documents on Arxiv database using Distributed Bag of Words. Qualified clusters was generated with 0.56 ARI and 0.59 Adjusted Mutual Information.	Number of clusters with data distribution is not explained. Time to create bag of words is not listed.
Noor S. Sagheer *et al* 2020 [21]	Partition clustering K-Means with Canopy preclustering	Big data- Health insurance	The work applied serial, Parallel modes. Parallel mode for K-means with canopy took less time 1.048 hour on insurance data.	The cluster quality is not proved.
Shan-shan Li *et al* 2020 [32]	Preprocessing: Replacement of missing values, Normalization. Density based DBSCAN with cover tree.	Heterogeneous applications	Cover tree data structure ease the process in finding neighbor similarity. Speeds up the process by having least time 19.68 seconds on benchmark datasets.	Number of clusters and the quality of the cluster is not specified.
Joelson Antonio dos Santos *et al* 2021[16]	Recursive Sampling with Density based DBSCAN, hierarchical Agglomerative.	Heterogeneous applications.	Sampling technique with data bubbles ease the analysis. Highest ARI 0.881 on benchmark datasets.	Memory utilization, processing time is not proved.

Table III, summarizes the unsupervised association rule mining with Apriori and FP-Growth algorithms with the applications supermarket, education, library, supply chain in company and some are implemented in sample transaction dataset.

TABLE III. . Summary of Related works on Unsupervised Learning method Association rule mining

Author,Year	Algorithm	Application	Merits	Demerits
Ashish Shah et al 2016 [4]	Improved Apriori, Hash function	Sample Transaction	Saves time and space with hash function. Number of scans to generate frequent itemset is reduced.	No. of rules, time to generate itemsets with, without hash function , is not given.
Andi et al 2018 [1]	FP-Growth	Library	Analyzed 1- frequent itemset, 2- frequent itemset and 3-frequent itemset. Generate 21 rules with confidence 50%.	Variant support, confidence values is not tried to get more number of itemsets and rules.
Le Hoang Son et al 2018 [18]	Apriori with Animal Migration Optimization	Supermarket	Produced compressed number of rules. Assessed with variant support and	Other than support, confidence values lift, leverage is not

			confidence values. 3120 rules were generated within 312 sec., 41.2 MB memory for the support, confidence values 0.1, 01 respectively.	included to have compact rules.
Dawei Dong et al 2019 [8]	FP-Growth with Improved Ant Lion optimization.	Education	Generate 170 rules.	Variant combination of support, confidence values is not tried.
Xu Hongfei et al 2019 [40]	Improved Apriori with undirected graph.	Sample Transaction	Instead of candidate set generation, undirected graph was used that reduce number of scans in generating frequent itemsets.	Association rule generation is not focused.
Huan-Bin WANG et al 2021[12]	Apriori with Parallelism.	Sample Transaction	Parallelism method leads to scan the transaction database only twice.	The work ends with frequent itemset but doesn't produce association rules.
Norulhidayah Isa et al 2021[22]	FP-Growth	Supply chain in company	With support 0.2, confidence 0.8 the frequent itemsets and rules were formed to get the pattern of purchasing.	Number of rules for the prescribed values is not specified.

III. CONCLUSION

Data mining categories supervised, unsupervised with its incorporated algorithms and preprocessing, text processing, procedures is very beneficial to have interesting and useful patterns in data. This survey discusses about recent works in variant applications with data mining methods classification, clustering, association rule mining elaborately to have in-depth knowledge.

It is observed in supervised learning, decision tree based algorithms Isolation forest, Random forest and C4.5 gave better results for the applications such as outlier analysis and for heterogeneous applications. SVM is better suits for text oriented applications. KNN is proved to be the best for biological analysis, outlier detection. Deep learning multilayer feed forward network applied on education analysis obtained accurate pattern. The Unsupervised clustering category has different performance evaluation measures based on its kind. The hybrid technique Partition based algorithm K-Means with probability based Expectation maximization on heterogeneous applications gave better result as it has less sum of squares and processing time and Fuzzy C-Means of the same category with PSO gave better F-Measure on heterogeneous applications. For customer segmentation, kernel based SVC provides least sum of squares thus proved cluster quality. The hybrid method

Density based DBSCAN with hierarchical proved its efficiency on heterogeneous dataset with high adjusted RAND index. For outlier detection, neural network based self organizing map with LOF factor produced good results with high cluster quality and outlier detection accuracy. The performance of unsupervised learning association rule mining states that combination of Apriori with Animal Migration Optimization proved to be the best for supermarket data, while considering number of scanning and processing time, Apriori with hashing, undirected graph, with parallelism gave better results, regarding FP-Growth, the algorithm with Improved Ant Lion optimization produced more number of rules. From the overall analysis it reveals, that all categories of algorithms if processed with suitable preprocessing and combined methods it gives efficient results on any type of applications.

REFERENCES

[1] Andi T, E Utami, "Association rule algorithm with FP growth for book search", 3rd Annual Applied Science and Engineering Conference (AASEC 2018), IOP Conf. Series: Materials Science and Engineering, pp. 1-6 , 2018.

[2] Anjuman Prabhat, Vikas Khullar, "Sentiment classification on Big Data using Naïve Bayes and Logistic Regression", International Conference on Computer Communication and Informatics (ICCCI -2017), IEEE publication, 2017.

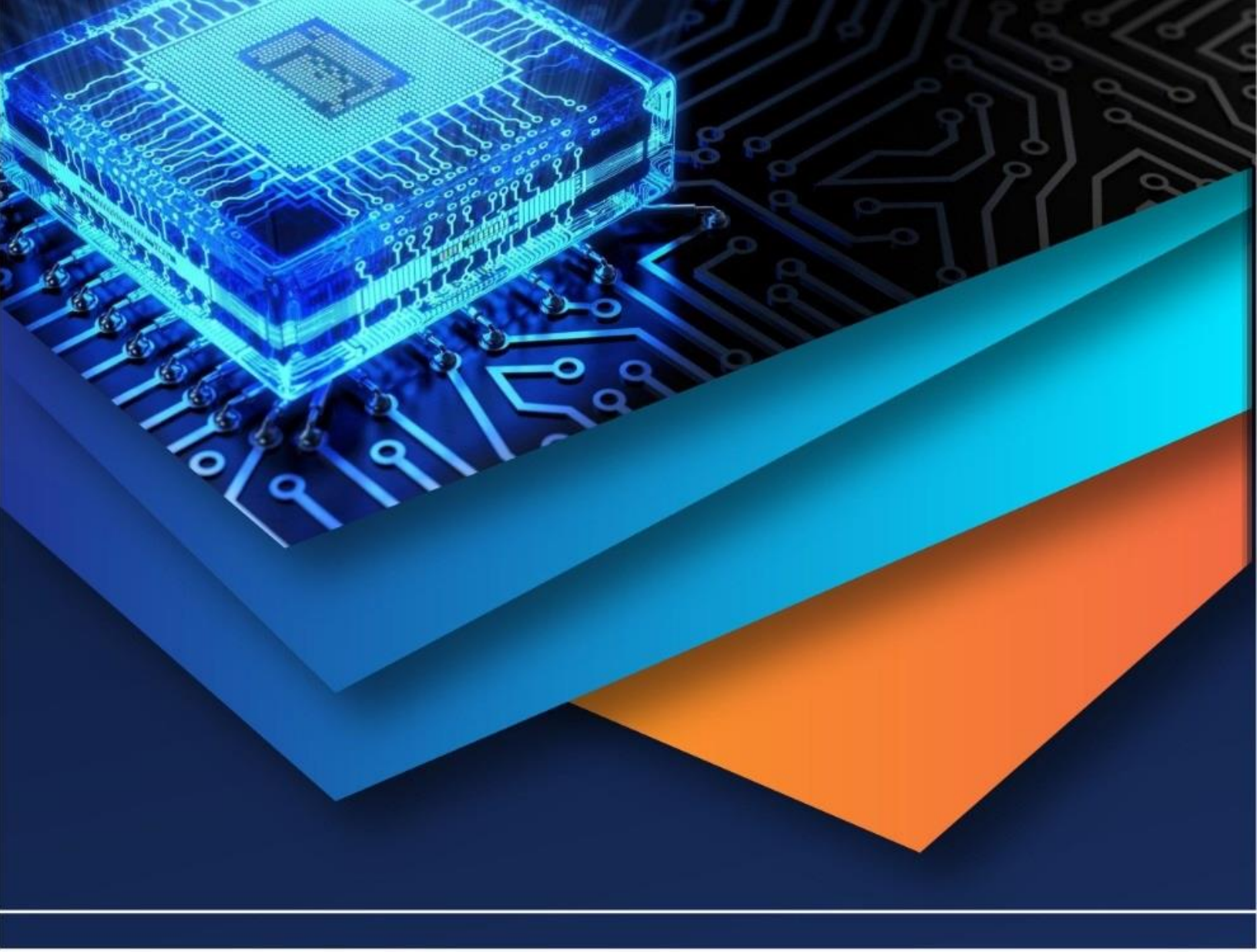
[3] Archana R. Panhalkar, Dharmpal D. Doye, "A novel approach to build accurate and diverse decision tree forest," Evolutionary Intelligence, Springer, pp.1-15, 2021.
[4] Ashish Shah, "Association Rule Mining with Modified Apriori Algorithm using Top down Approach", 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), IEEE Publication, pp. 747-752, 2016.
[5] Ayesha Anzer, Hadeel A. Tabaza, Jauhar Ali, "Predicting Academic Performance of Students in UAE Using Data Mining Techniques", International Conference on Advances in Computing and Communication Engineering (ICACCE-2018), pp. 179-183 , 2018.
[6] Chittem Leela Krishna, Dr. Poli Venkata Subba Reddy, "An Efficient Deep Neural Network Multilayer Perceptron Based Classifier in Healthcare System", Third International Conference On Computing And Communication Technologies(ICCCT'19), IEEE Publication, pp.1-6, 2019.
[7] Daniel T. Larose , Chantal D. Larose, "Data Mining and Predictive Analytics", Second Edition, Wiley Publishing, 2015.
[8] Dawei Dong, Zhiwei ye, Hubei University of Technology, Yu Cao, Shiwei Xie, Fengwen Wang, Wei Ming, " An Improved Association Rule Mining Algorithm Based on Ant Lion Optimizer Algorithm and FP-Growth", 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), IEEE Publication, pp. 458-463, 2019.
[9] Dingsheng Deng, DBSCAN Clustering Algorithm Based on Density", 7th International Forum on Electrical Engineering and Automation (IFEEA), IEEE Publication, pp.949-953, 2020.
[10] Farhan Ullah, Bofeng Zhang, Rehan Ullah Khan, Tae-Sun Chung, Muhammad Attique, Khalil Khan, Salim El Khediri, Sadeeq Jan, " Deep Edu: A Deep Neural Collaborative Filtering for Educational Services Recommendation", IEEE Access, pp. 110915-110928, 2020.
[11] He Xu, Lin Zhang , Peng Li, Feng Zhu, "Outlier detection algorithm based on k-nearest neighbors-local outlier factor ",Journal of Algorithms & Computational Technology, pp.1-12, 2021.
[12] Huan-Bin WANG, Yang-Jun GAO, " Research on parallelism of Apriori algorihtm in association rule mining", 10th International Conference of Information and Communication Technology (ICICT-2020), Elsevier, pp.641-647, 2021.
[13] Hussain Ahmad Madni; Zahid Anwar; Munam Ali Shah, "Data mining techniques and applications A decade review", 23rd International Conference on Automation and Computing (ICAC), IEEE publication, 2017.
[14] Iulia-Maria RADULESCU, Ciprian-Octavian TRUICA, Elena Simona APOSTOL, Alexandru BOICEA, Mariana MOCANU, Daniel Calin POPEANGA, Florin RADULESCU, "Density-based Text Clustering using Document Embeddings", 36th International Business Information Management Association (IBIMA), Sustainable Economic Development and Advancing Education Excellence in the era of Global Pandemic, pp. 6222-6230, 2020.
[15] Jain .N and V. Srivastava, "Data Mining Techniques: a Survey Paper," IJRET Int. J. Res. Eng. Technol., vol. 2, no. 11, pp. 116– 119, 2013.
[16] Joelson Antonio dos Santos, Talat Iqbal Syed, Murilo C. Naldi, Ricardo J. G. B. Campello, Joerg Sander, " Hierarchical Density-Based Clustering Using MapReduce", IEEE Transactions on Big Data, Vol. 7, NO. 1, pp. 102-114, 2021.
[17] Lalit Mohan, Janmejay Pant, Priyanka Suyal, Arvind Kumar, " Support Vector Machine Accuracy Improvement with Classification", 12th International Conference on Computational Intelligence and Communication Networks, IEEE, pp. 477-481, 2020.
[18] Le Hoang Son, Francisco Chiclana, Raghavendra Kumar, Mamta Mittal, Manju Khari, Jyotir Moy Chatterjee , Sung Wook Baik, "ARM-AMO: An Efficient Association Rule Mining Algorithm Based on Animal Migration Optimization", Knowledge-Based Systems, Elsevier, Volume 154, pp. 68-80, 2018.
[19] Najat Ali, Daniel Neagu, Paul Trundle, "Evaluation of k-nearest neighbour classifer performance for heterogeneous data sets", Springer nature, 2019.
[20] Niko Reunanen, Tomi Räty, Timo Lintonen, "Automatic optimization of outlier detection ensembles using a limited number of outlier examples", International Journal of Data Science and Analytics, Springer Publication, pp.377-394, 2020.
[21] Noor S. Sagheer, Suhad A.Yousif, "Canopy with k-means Clustering Algorithm for Big Data Analytics", Fourth International Conference of Mathematical Sciences (ICMS 2020), AIP publication, pp. 070006-1- 070006-4, 2020.
[22] Norulhidayah Isa, Siti Khadijah Neddy, Norizan Mohamed, "Association Rule Mining using FP-Growth Algorithm to Prevent Maverick Buying", IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE Publication, pp. 77-81, 2021.
[23] Parteek Bhatia, "Data Mining and Data Warehousing: Principles and Practical Techniques", Cambridge University Press Publishing, 2019.
[24] Pedro albuquerque, Solange alfinito, Claudio v. Torres, "Support Vector Clustering for Customer Segmentation on Mobile TV Service", Communications in Statistics - Simulation and Computation, pp. 1453-1464, 2015.
[25] Ping Yang, Dan Wang, Zhuojun Wei, Xiaolin Du, Tong Li, "An Outlier Detection Approach based on Improved Self-Organizing Feature Map Clustering Algorithm", AI-Driven Big Data Processing: Theory, Methodology, and Applications, IEEE Publication, pp. 1-13, 2019.
[26] Praveen .P, "An Efficient Linkage Criterion for Creating Clusters in Hierarchical Method", International Journal of Future Generation Communication and Networking Vol. 12, No. 5, pp. 294- 300, 2019.
[27] Raed Alsini , Abdullah Almakrab , Ahmed Ibrahim , Xiaogang Ma, "Improving the outlier detection method in concrete mix design by combining the isolation forest and local outlier factor", Construction and Building Materials, Elsevier, pp. 1-7, 2020.
[28] Raja Kishor, D , N. B. Venkateswarlu, "Hybridization of Expectation-Maximization and K-Means Algorithms for Better Clustering Performance", Cybernetics And Information Technologies, Volume 16, No 2, pp. 16-34, 2016.
[29] Rishabh Ahuja, Arun Solanki, Anand Nayyar, "Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor," 9th International Conference on Cloud Computing, Data Science & Engineering, IEEE publication, pp. 263-268, 2019.
[30] Sanditya Hardaya I. B. N., Arian Dhini, Isti Surjandari, "Application of text mining for classification of community complaints and proposals", 3rd International Conference on Science in Information Technology (ICSITech), IEEE Publication, pp.144-149, 2017.
[31] Saptarshi Sengupta, Sanchita Basak, Richard Alan Peters II, "Data Clustering using a Hybrid of Fuzzy C-Means and Quantum-behaved Particle Swarm Optimization", IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), IEEE publication, pp.137-142, 2018.
[32] Shan-shan Li, " An Improved DBSCAN Algorithm Based on the Neighbor Similarity and Fast Nearest Neighbor Query", Artificial Intelligence in Parallel and Distributed Computing, pp.47468-47476, 2020.
[33] Shu-Hsien Liao , Pei-Hui Chu, Pei-Yuan Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011", Expert Systems with Applications, Elsevier, pp. 11303-11311, 2012.

[34] Sweetline Priya .E, Anandhan .K, "An Overview of Data Mining - A Survey Paper", International Journal of Modern Computer Science (IJMCS), Volume 6, Issue 1, pp.19-21, 2018.
[35] Uma Maheswari .B, Sujatha .R, "Introduction to Data Science: Practical Approach with R and Python", Wiley Publishing, 2021.
[36] Valliammai V , Suruthi Selvi S , Akshaya Sibi S P , Nidharshna M , Lavanya U, "Survey on Different Data Mining Algorithm for Prediction", International Advanced Research Journal in Science, Engineering and Technology Vol. 8, Issue 8, pp.442-446, August 2021.
[37] Vipin Kumar, Pang-Ning Tan Michael Steinbach, "Introduction to Data Mining", First edison, Pearson Publication, 2016.
[38] Wenxian Feng, Chenkai Ma, Guozhang Zhao, Rui Zhang," FSRF:An Improved Random Forest for Classification", IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), pp.173-178, 2020.
[39] Xiaoyu Luo, "Efficient English text classification using selected Machine Learning Techniques," Alexandria Engineering Journal, Elsevier, pp. 3401-3409, 2021.
[40] Xu Hongfei, Xuesong Liang, Wei Cui, Wei Liu, "Research on an Improved Association Rule Mining Algorithm", IEEE International Conference on Power Data Science (ICPDS), IEEE publication, pp. 37-42, 2019.

Author Profile

Dr. S. Sukumaran, working as Associate Professor and Head of Department, Department of Computer science (Aided) in Erode Arts and Science College, Erode, Tamilnadu, India. He is a member of Board of studies in various Autonomous colleges and universities. In his 36 years of teaching experience, he has supervised more than 55 M.Phil Research Works, guided 24Ph.D research works and still continuing. He has presented, published around 80 research papers in National, International Conferences and Peer Reviewed Journals. His area of research interest includes Digital Image Processing and Data mining.

S.Sivasankari, has completed B.Sc (Computer science) in affiliated college of Madurai Kamaraj University, M.C.A at Bharathiar University, Coimbatore in distance education. She has awarded M.Phil in Data Mining from Bharathiar University, Coimbatore. At present, she is continuing her doctorate research work (Part time) in Data Mining at Department of Computer science (Aided), Erode Arts and Science College, Erode, Tamilnadu, India.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY