International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

A Systematic Methodology on Class Imbalanced Problems involved in the Classification of Real-World Datasets

K. Santhi , A Rama Mohan Reddy

Abstract : Current generation real-world data sets processed through machine learning are imbalanced by nature. This imbalanced data enables the researchers with a challenging scenario in the context of perdition for both the machine learning and data mining algorithms. It is observed from the past research studies most of the imbalanced data sets consists of the major classes and minor classes and the major class leads the minor class. Several standards and hybrid prediction algorithms are proposed in various application domains but in most of the real-time data sets analyzed in the studies are imbalanced by nature thereby affecting the accuracy of the prediction. This paper presents a systematic survey of the past research studies to analyze intrinsic data characteristics and techniques utilized for handling class-imbalanced data. In addition, this study reveals the research gaps, trends and patterns in existing studies and discusses briefly on future research directions.

Keywords: Data intrinsic problem, imbalanced data sets, Machine learning, prediction algorithms.

I. INTRODUCTION

Most of the classification problems suffer from accuracy and performance due to the imbalanced datasets as these are not categorized equally. Equal class distribution is assumed by various machine learning and deep learning algorithms during the process of prediction. But in the context of the real world scenario, it is not possible because the data in this context will be skewed in nature as one class of the data is most frequently repeated than the other class of the data. As most of the machine learning algorithms give priority to the major classes this could be addressed as a major problem while designing automated and smart systems. Learning from imbalanced data [1] plays a major role in the development of automated prediction systems.

Recently many researchers have drawn their significant attention towards the classification of imbalanced datasets that helps during the process of improving the performance of the prediction algorithms [2]. Initially, in 1993 Anand et al.[3] addressed the classification problem on proposing a neural network based algorithm to handle the imbalanced data set.

Manuscript published on 30 September 2019
* Correspondence Author
K Santhi*, Research Scholar , Department of Computer Science and Engineering, S V University College of Engineering, Tirupati, India.
Email: santhi.kuraganti@gmail.com
A Rama Mohan Reddy, Department of Computer Science and Engineering, S V University College of Engineering , Tirupati, India.
Email: ramamohansvu@yahoo.com

Further, in later years, many researchers have developed significant and efficient algorithms that mainly focus on the modification of the existing classifiers and utilization of the pre-processing techniques to balance the imbalanced data set. Literature studies in [3] indicate that conventional machine Learning algorithms always assume that the datasets consist of minute misclassification error rate as well as they are well balanced. Data intrinsic problem in the context of the classification problem arises due to the variable number of samples in each class. If there are a large number of samples in one class it is notated as majority class and if there are a small number of samples it is notated as the minor class. These type of skewed distributed datasets indicates class imbalance problems.

With the engrossed attention of many researchers towards the class imbalance problems, it is observed from many research studies that a testing sample data set is allocated to the majority class if the available the sample size of the each associated class is imbalanced. In this context, an appropriate resampling method is adopted along with a powerful classifier to handle this problem. Further, it is observed in a few research studies that the multi-classifier algorithms outperform the range of the single classifier algorithm during the implementation of the vast range of algorithms. Real-world datasets like medical datasets, insurance datasets, banking datasets, network datasets that include unreliable telephonic sources, data sets used in information filtering and retrieval tasks, etc are generally affected by the class imbalance problems. A detailed survey of the class imbalanced problem in datasets is presented by Japkowicz et al.[1] in which classification of imbalanced data, as well as the range of the damage, stumble upon the prediction mechanism while imbalanced data set is trained with a classifier algorithm. It is observed in the study that, this survey does not include several techniques, algorithms, approaches, and metrics that handle the class imbalance problem in the dataset. Further a vast range of research is carried out in the context of developing various algorithms and approaches to handle the data intrinsic problem. To the finest of our knowledge, there is no such systematic survey with the inclusion of details regarding the variety of algorithms and approaches that address the data intrinsic problems.

This article presents a systematic study of significant approaches and algorithms developed addressing the class imbalance problem during the past decade. Detailed pros and cons of these algorithms are analyzed thoroughly to identify the research gap that could enhance the priority of research in improving the accuracy of the prediction
Algorithms. The contribution of this study is to focus on the identification of the preliminary objectives regarding the effect of the imbalanced datasets in classification and indicate various techniques used in this context.

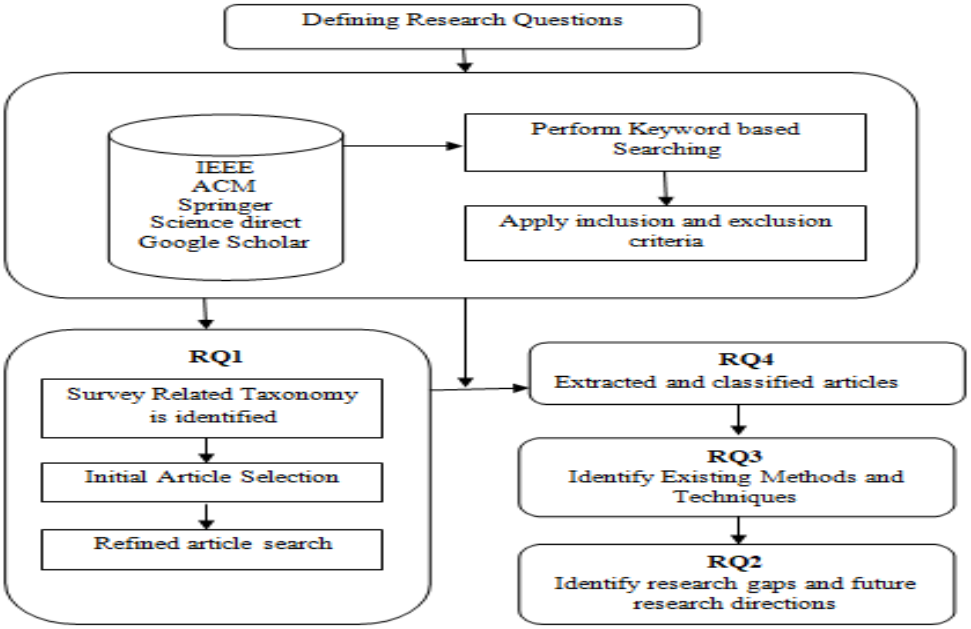

Figure 1 Research Methodology

II. RESEARCH METHODOLOGY

Systematic Literature Review (SLR) includes the evaluation and interpretation of the existing research studies which is related to the specific topic, research question and the observable fact of interest [4]. The main objective of the research methodology is to provide an unbiased as well as the systematic review that could be repeatable and auditable by the next generation researchers. This article adopts a four-phase procedure for the systematic study as shown in Figure. 1 which includes the process of formulating the research questions along with the motivation behind it, searching and extraction of the relevant data as well as assessing the quality of the data [4, 5]. This process enables the next generation researchers with a systematic understanding of the existing approach, techniques, solutions, implications and future directions relating to the research of data mining and machine learning. Further, it includes the algorithm and approaches along with their constraints to understand the purpose of research.

A. Defining Research Questions

To initiate the process of SLR initially the research questions relating to the survey are defined along with the motivations behind it. Table 1 enlists the research questions along with their corresponding motivations

B. Searching the relevant data

The intention of the systematic study is to discover, compare and classify the existing research studies within the class imbalance problems using a systematic procedure. Popular scientific databases like IEEE, SPRINGER, ACM, SCIENCE DIRECT and Google scholar libraries are utilized for the searching process using the following search strings in different combinations In the initial cases during the study using the above search string we have identified around 300 research papers in specific to the class imbalanced problems that are published in the past decade. Table 2 indicated different online mechanisms utilized for the process of searching relevant studies.

Table 1: Formulation of Research Questions

Motivation	*Research Question*
Gain insights relating to the Class imbalance problem	RQ1: What are the primary research motivations behind the data intrinsic problem?
Attain Knowledge regarding the status of research and future directions in the field of intrinsic data analysis	RQ2: What is the current status of research and future scope for data intrinsic problem?
Solutions that are available to balance the imbalanced datasets in classification	RQ3: What are the existing algorithms and approaches that are applied for data intrinsic problem?
Identify the necessity of balancing the unbalanced datasets and their effect on the accuracy of prediction.	RQ4: Why class imbalanced problems are popular?

Class imbalance problem **OR** Data intrinsic problem for imbalanced dataset classification
AND
. Minority classes **OR** Majority classes **OR** Skewed distribution
AND
Oversampling Methods **OR** under sampling methods **OR** Sampling methods
AND
. Data specific method **OR** Algorithm Specific Methods **OR** Hybrid Mechanisms
AND
Systematic Study **OR** SLR **OR** Mapping Study **OR** Review

Table 2: Various Scientific databases considered for SLR

[1] S. No	[2] Database	[3] No. Of Papers
[4] 1	[5] IEEE	[6] 178
[7] 2	[8] ACM	[9] 10
[10] 3	[11] SPRINGER	[12] 42
[13] 4	[14] SCIENCE DIRECT	[15] 45
[16] 5	[17] GOOGLE SCHOLAR	[18] 25
[19] Total		[20] 300

B.1. Preliminary selection

In the preliminary phase, of the selection 300 related articles are extracted from the scientific databases using the search strings in which articles are scrutinized based on the relevancy of the title towards the problem statement and further the remaining are ignored. Additionally, the articles from the thesis, book chapters, short papers and papers communicated in non-English language are ignored from the study. The inclusion and exclusion criterion of the articles is made by analyzing the title, abstract and conclusion of the articles. Such that filtering of the appropriate articles is accomplished with relevance to the class imbalance problem.

Title: (Data centric techniques **OR** Data Level Approaches for class Imbalance **OR** Data Intrinsic Problems in classification **OR** Algorithm Centric techniques **OR** Neural Networks **OR** Convolution Neural Networks **OR** FUZZY techniques **OR** Hybrid Mechanisms **OR** Ensemble learning techniques **OR** Swarm Intelligent Techniques)
AND
Abstract: (Data sampling mechanisms **OR** Techniques and approaches of class imbalances **OR** under sampling **OR** Over sampling techniques **OR** Decision Tree **OR** SVM)

B.2. Selecting the articles based on the Implementation details and Data sets Utilized

In this phase, we further analyses the quality of the articles based on the implementation details furnished in the article, such that the article is thoroughly examined to identify the algorithm utilized to resolve the data intrinsic problem in classification along with its implantation with real-world dataset are considered for the review process. Further the articles with detailed description of algorithm are considered based on the novelty of the technique. Based on the opinions and suggestions of the experts with experience in the
Process of conducting systematic review in various domains, search strings are modified to be more focused towards the topic and furnished above. Feedback from a team of research experts is considered to enhance the process of SLR. The detailed discussion of various algorithms, techniques and approaches regarding the class imbalance problems is provided in section 3.

III. VARIOUS TECHNIQUES IN THE CLASSIFICATION OF DATA INTRINSIC AND CLASS IMBALANCE PROBLEMS

The class imbalance problems in selected literature studies are broadly addressed in two different levels that include data and algorithm centric mechanisms, and Ensemble or hybrid algorithms. In the context of data centric techniques the data set to be trained is modified to balance the class distribution such that they utilize pre-processing techniques to eliminate skewed distribution of the data. Algorithm centric techniques concentrate on modification of existing learning algorithms that elevate the biased performance towards the major classes and adapt them for mining the skewed distribution. In the context of the hybrid mechanisms various classifiers are trained to combine the data into a single level class.

A. Data and Algorithm Centric Mechanisms:

Li et al.in [6] developed a novel method called oversampling by using support degree and with this method, a new minority class samples are generated by the selected minority samples. It also helps to identify the boundary class samples to generate the new boundary samples along with their neighbours. To perform the classification on the class samples of the original dataset, a synthetic sample has to be added. To overcome the problems of re-sampling traditional methods like information loss, severe randomness, and subjective interference can be done by introducing under-sampling techniques that are proposed by Li et al. [7].To get solved the problem of imbalanced class data Dang et al. [8] develops a new technique called SPY. In this, a majority class can be represented as spy samples and these majority class samples label in the training dataset can be easily changed into minority class samples label.

To attain accuracy on the balanced data Koto et al. [9] improves SMOTE in three different ways and named those improvements as Selected-SMOTE, SMOTE-Cosine, and SMOTE-Out He performed these three improvements on different eight datasets and collected the results. Then these results are compared with the SMOTE results to know the improvements done on new methods Oktavino et al. [10] develops a high performance automatic algorithm to extract information from e-commerce website to process the text as well as balance the datasets by using SMOTE.

Based on the previous studies it is observed that under-sampling techniques can have data loss problem. So, two methods are developed by Zhang et al. [11] to find out the majority classes which are not used by under sampling. The two methods such as K-means and random sampling are used to predict minority classes and to prevent information loss from majority class. In a minority class, synthetic samples are produced by developing a novel technique called over-sampling by an author named Wang et al. [12]. To balance unbalanced data samples, a graph called K-NN is drawn by taking positive classes. To forecasting the faults of software data while improving the unbalanced dataset to balanced data an author named Fu et al. [13] developed an algorithm for Principal Component Analysis (PCA) and under-sampling method. Further to solve imbalanced problems of class while finding the defects in software more concentration has to be maintained to improve the performance of AUC. So, an author Ma et al. [14] proposes a new technique called RusTri (Random under sampling Tri-training).

To enhance the performance of classification and accuracy and also to examine the methods of under-sampling while solving the class distribution imbalance problems a cluster-based technique of sampling was developed by Wu et al. [15] to select the data as representative to the entire data set from the datasets. To balance geometric feature extraction of the imbalanced dataset a Majumder et al. [16] develops new algorithms called SMOTE and Adaptive Synthetic Sampling (ADASYN). An imbalance dataset of the class can be solved by proposing an under-sampling method by an author named Shi et al. [17] and to get trained the datasets of re-sampled an SVM classifier is used. On the basis of missing data, an author named M Faraajzadeh et al. [18] develops new sampling techniques to rectify the defects of imbalanced problems of class as a reason of k-NN and Expectation Maximization.

As a reason of oversampling, to solve the imbalance problem of class which uses Support Vector Data Description (SVDD) can be done by a non-synthetic oversampling algorithm which was proposed by Ghazikhani et al. [19]. The non-synthetic oversampling algorithm has two steps. Firstly, on lower class a Support Vector Data Description (SVDD) is performed then oversampling of the support vectors is done. Hu et al.[20] to identify the class imbalance problem a modification is done on SMOTE algorithm to get the better performance. Li et al. [22] proposed new method called Random-SMOTE based on over-sampling to solve the class imbalance problems. An author named Liu et al. [21] improvises the method called SMOTE by an over-sampling method i.e., K-Means algorithm and under-sampling method i.e., RUS for data balancing. Cost Minimization Oriented SMOTE (CMO-SMOTE) was proposed by Zhou et al. [23] for adjusting the boundaries of class minority datasets.

To balance the class datasets high and low accuracy, Bhowan et al. [24] suggest the algorithm called genetic programming (GP) that effectively develops high-performance classifiers. The imbalanced problem can be solved when an author named uangthong et al. [25] develops Rotation Forest-J48 with SMOTE. To reduce the sampling data set size of each class an algorithm of under sampling using fuzzy logic was developed by Wong et al. [26].

A novel algorithm was proposed by Padmaja et al. [27] to detect the fraud, the author uses k-Reverse NN (kRNN) algorithm to remove extreme outliers. Zhang et al. [28] developed an algorithm to balance the dataset classification of unbalanced classification which integrates that classification and ensemble learning algorithms. The author also develops a hybrid algorithm that combines 3 methods which are random feature selection, under sampling and bootstrap re-sampling. This method processes the trained data sets using base classifiers. Fan et al. [29] explain that if the boundaries of each class datasets are very balanced then the re-sampling methods are not necessary.

Support Vector Machines gives effective performance when compared with resampling methods like RUS and SMOTE. To identify the streaming of imbalanced data, Chen et al. [30] developed selectively recursive approach (SERA). To rebalance the data set in [31, 32] an RBF classifier is used.To balance the unbalanced data samples Hosseinzadeh et al. [33] develops classifiers which perform effectively using SMOTE oversampling and also explains the methods of RF and Fuzzy C-means. To reduce imbalanced data a distance of boundaries has to be adjusted and found by using Mahalanobis but not by Euclidian distances and these methods was proposed by an author in Simple Hybrid Sampling Approaches (SHSA).

Shin et al. [34] modifies the distances of boundaries by applying different cost sampling approaches to different classes. This problem was overcome by an author Cohen et al. [35], he examines the Support vector machines of one class by using different class examples and he improves the performance by the transformation of the conformal kernel. With these, the imbalances of sampling can be improved with the help of learning algorithms of one class with other class ignoring and by adjusting the [36] asymmetrical margins of rare positive cases.

As per the imbalanced distribution of data, the adjustment of boundaries can be done with changing the values of kernel matrix and this idea was by Wu et al. [37]. An efficient base classifier [38] can be identified by a mono meta-classifier called stacking. Stacking bagging cost and performance can be improved with the help of bagging called base classifier predictions and all these were proposed by the author named Phua et al. This author also developed a new algorithm with of resampling methods to build new decision trees by using an algorithm called consolidated tree construction (CTC).

Table 3: Various Data and Algorithm Centric approaches

Author Name	Algorithm	Technique	Dataset Utilized	Metric
Mustafa et al.[39]	KNN based fuzzy-rough set	Maximum distance based SMOTE	UCI repository	Accuracy
Dang et al. [8]	SPY	SVM, K-NN, and Radial Function(RF)	UCI repository	Sensitivity, G-Mean
Li et al. [6]	C4.5	SDSMOTE	UCI Repository, JM1 is from NASA standard data-sets	F-measure , G-Mean ,AUC
Li et al.[7]	Support vector Machine (SVM)	Sampling Strategy based on SOM	UCI repository	F-measure, Precision, Recall
Koto et al. [9]	SVM	SMOTE Out, SMOTE-Cosine, Selected-SMOTE	UCI repository	F-measure
Oktavino et al. [10]	SVM, NB, DT	SMOTE	Indonesia e-commerce	Accuracy
Zhang et al. [11]	NB	Cluster-based Majority Under-Sampling	UCI repository	Recall, Precision, F-measure, G-Mean
Wang et al. [12]	SVM	ROS	UCI repository	Sensitivity, Specificity, Accuracy, G-Mean
Wu et al. [15]	DT	Hybrid-Sampling Technique	KEEL repository	AUC, F-measure
Majumder et al. [16]	GMM	SMOTE, ADASYN	Shoulder Pain Expression Archive database	Precision, Recal
Shi et al. [17]	SVM	HIOSVM	UCI repository	F-measure, G-Mean, AUC
Farajzadeh et al. [18]	Adaboost.M1, Bagging	SMOTE	CWRU bearing data center	Normalised Popt
Ghazikhani et al. [19]	KNN	SVM	UCI repository	F-measure, Recall, Precision
Hu et al. [20]	C4.5, AdaBoost	MSMOTE	UCI repository	Precision, Recall, F-measure
Li et al. [21]	NB	SMOTE with K-means	IPTV datset	Accuracy, F-measure, G-Mean, Precision, Recall
Li et al. [22]	LR	Random- SMOTE	UCI repository	Accuracy
Zhou et al. [23]	NB	CMOSMOTE	KEEL repository	Recall, Precision, G-Mean, F-measure
Bhowan et al. [24]	GP	Cost Adjustment	UCI repository	Accuracy
Ruangthong et al. [25]	RF, J48	SMOTE	UCI repository	Sensitivity, Specificity, Accuracy
Wong et al. [26]	CHC algorithm	SVM	UCI repository	F-measure, AUC
Padmaja et al. [27]	DT, KNN, NB, RBF	Majority filter-based minority prediction	UCI repository	Accuracy, Recall, Precision, F-measure

Zhang et al. [28]	DT	Hybrid method	Tans hotel	Precision, Recall, Accuracy, F-measure, AUC, G-Mean
Fan et al. [29]	SVM	Re-sampling using the boundary ratio	UCI repository	Recall, Precision, F-measure, G-mean
Chen et al. [30]	BBagging	SERA	KDD cup 1999 network intrusion dataset	Precision, Recall, F-measure, G-Mean
Hosseinzadeh et al. [33]	C4.5	Ensemble method (fuzzy C-means, Rotation Forest)	KEEL repository	AUC
Shin et al. [34]	SVM	Different Cost	Direct Marketing Education	ROC
Cohen et al. [35]	one-class SVM	kernel transformation	nosocomial dataset	Accuracy, Sensitivity, Specificity

B. Hybrid Mechanisms addressing Imbalanced class problem

Trained models performance will be improved in imbalanced datasets by a new ensemble method which was proposed by Wei et al. [40] called Balanced Boost. Balance Cascade and Easy Ensemble methods were proposed by Liu et al. [41] for the learning of class-imbalance. All these ensembles or hybrid algorithms uses the examples of class which was not used in under sampling. With these, all the weak learners from the subsets of multiple major classes will ensemble with each other to get the final results. Ceballes et al. [42] combine sub sampling, Boosting and oversampling to develop a new hybrid method called Swarm Boost to get samples. To get samples from the existing training sets a Jiang et al. develops three methods such as SBNC-oversampling, SBNC-hybrid and sampled Bayesian network classifiers (SBNC)-under sampling. For solving class imbalance problems an author called Oliveira et. al.[43] develops Iterative Classifier Selection Bagging (ICS-Bagging) to generate an ensemble of the classifier. Ruangthong et al. [44] predict Bank customers term deposit probability while class imbalance problem was solving, a hybrid ensemble technique was developed by incorporating the techniques in AdaBoost.M2 and adopt SMOTE. The forecasting performance was increased by new ensemble technique AdaBoost with SMOTE, which was proposed by Huang et al. [46]

Pal et al. [47] for efficient clustering a new model was developed called BoostedGMM by using SMOTE Cluster and Samplings of SMOTE. A New Clustering based Subset Ensemble Learning method was developed by Hu et al. [48] for class imbalances. To solve the class imbalance problems Mustafa et al. [49] proposed a novel hybrid technique called "Distribution based MultiBoost (DBMB)" by using machine learning techniques. To identify the claims called credit card churn prediction and automobile insurances between the two datasets a new method was proposed by Sundar kumar et al. [50] called one-class support vector machine (OCSVM).

Krawczyk et al. [51] identify different unbalanced objects in the category of malignant while the samples are examining. By this author compare the performances of classifiers like Pruned Under-Sampling Balanced Ensemble (PUSBE), Boosted Support Vector Machine, Hybrid Cost-Sensitive Ensemble (HCSE) and SMOTEBoost. The problem of class imbalance can be solved by GAO Et Al. [55] after using RUSBoost learning method. Yongqing et al. [52] add a new adaptive algorithm to the SMOTE and then it improvises the old SMOTE into Adaptive SMOTE (ASMOTE). This new algorithm solves the problem of class imbalance by modifying the nearest neighbours. Wu et al. [53] designed a model by considering improved AdaBoost and SMOTE to predict the chain of e-commerce customers. To improve the class classification and prediction of class imbalance a Thai et al. [54] uses cost sensitivity learning and sampling methods.

Clustered Knowledge Management Development Framework (CKMD) was developed by El et al. [56] for data imbalance for the efficient retrieval of knowledge and also increases the knowledge discovery performance. For the classification of Gene in imbalanced data, a new technique was by developed by Soltani et al. [57]. Zughrat et al. [58] proposed a method for the classification of imbalanced data called "iterative fuzzy support vector machine algorithm (IFSVM)" with bootstrapping-based oversampling and under-sampling.

Table 4. Various Hybrid approaches

Author Name	Algorithm	Technique	Dataset Utilized	Metrics
Liu et al. [41]	EasyEnsemble	EasyEnsemble	UCI repository	ROC
Liu et al. [45]	AdaBoost	Cost-sensitive learning	Jiangsu Telecom	F-measure
Oliveira et al. [43]	Iterative Classifier with Bagging	SMOTE-ICS-Bagging	KEEL repository	ROC
Ruangthong et al. [44]	BN, DT, J48	AdaBoost.M2, SMOTE	UCI repository	Sensitivity, Specificity, Accuracy
Huang et al. [46]	AdaBoost	AdaBoost, SMOTE	SKEMPI	Precision, Recall, Specificity, Accuracy, F-measure
Sandhan et al. [59]	SVM, LR, k-NN, Gaussian classifier	Hybrid sampling, Bootstrapping	structural classification of proteins	ROC
Winata et al. [60]	DT (J48), NB, SMO	Adaptive boosting, bagging	LAPOR dataset	Hamming loss, accuracy, F-measure
Dwiyanti et al. [61]	C4.5	RUSBoost	PT. Telkom Indonesia	F-measure
Galar et al. [62]	C4.5	EUSBoost	ROC	ROC
Fernandez et al. [63]	RF	RUS, ROS, SMOTE	Big Data UCI repository	G-Mean
Qian et al. [64]	NB	Resampling ensemble algorithm	UCI repository	Precision, Recall, G-Mean, F-measure
Tran et al. [65]	One-Class Classification, AdaBoost	RABOC	BioSecure DS2, XM2VTS	Half Total Error Rate
Barandela et al. [66]	1-NN(Multi-Layer Perceptron)	ensemble learning	UCI repository	Accuracy
Jiang et al. [68]	C4.5	GA-based SMOTE	KEEL repository	F-measure, G-Mean
Guo et al. [67]	DataBoost-IM	ensemble learning	UCI repository	F-measures, G-mean, accuracy
Sundarkumar et al. [50]	DT, SVM, LR, PNN	VM based Undersampling	Automobile Insurance fraud and Credit card customer churn dataset	Sensitivity, Specificity, Accuracy

III. ANALYSING THE QUALITY OF DATA

This section discuss about the results of our literature study conducted on the class imbalance problems on addressing the various research questions shown in Table 1.

RQ1: What are the primary research motivations behind the data intrinsic problem?

- The main research motivation regarding the class imbalance problem is discussed in section 1 and 2.

RQ4: Why class imbalanced problems are popular?

- The rate of attention gained by the researchers on the class imbalance problems are justified by analysing the following points.

A. Identify number of publication in specific to the class imbalance problems:

The rate of publication of the articles addressing the class imbalance and data intrinsic problems in scientific databases from a decade is shown in Figure 2. These quantified results of the publication count specify that the rate of publication in IEEE databases is far greater than the result of any other database.

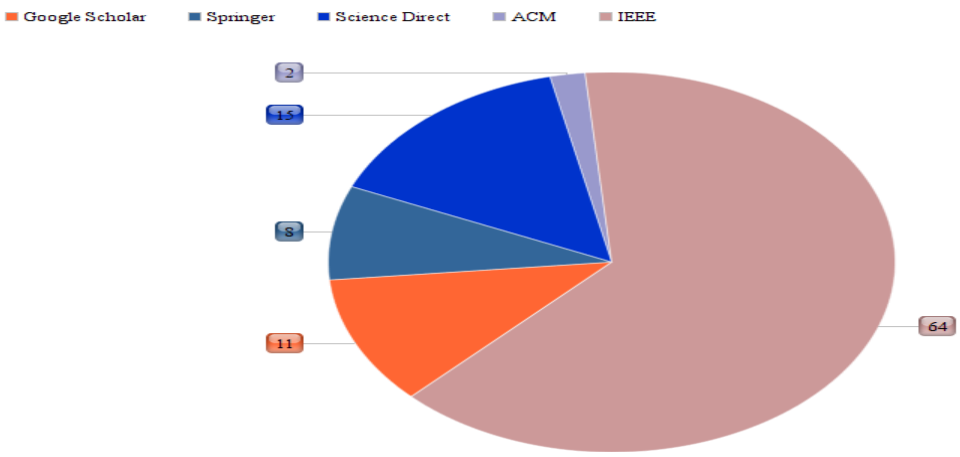

Figure 2 . Distribution of articles from relevant scientific databases

B. Analyse the growth factor of the article publication to the class imbalance problems:

The number of research publications on the data intrinsic problem and year of the publication is depicted in Figure 3. It is observed that there is year wise significant growth in the number of publication which strengthens the phrase that the class imbalance problem and data intrinsic problems are gradually gaining the attention of researchers every year. Form the analysis it is identified that most of the publications are from the major conference proceedings such that among 300 collected article around 64% of articles are related to conference proceedings. Furthermore, 52 journals are published in major journals that include Neuro computing, Expert and intelligent systems, Journal of Soft computing and knowledge based systems.

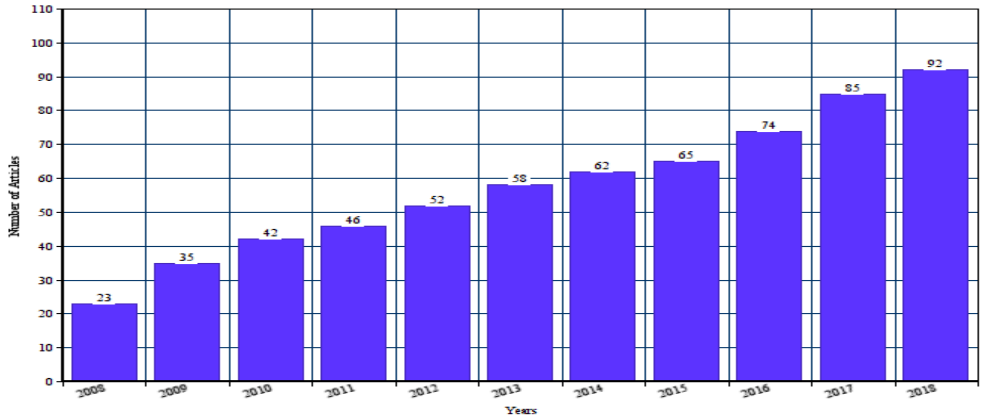

Figure 3. Rate of increment in the articles published every year

RQ3: What are the existing algorithms and approaches that are applied for data intrinsic problem?

Based on the extensive literature study conducted in Section.3 the data intrinsic problems are broadly classified into data and algorithm centric techniques and ensemble and hybrid techniques. It is observed that most of the articles i.e., around 72% of articles belongs to data and algorithm centric techniques addressing class imbalance problem and 28% of the articles have addressed hybrid mechanisms.

RQ2: What is the current status of research and future scope for data intrinsic problem?

The following section 5 elaborates in detail about unaddressed research challenges and open research directions in the application of various algorithms that handles class imbalance problems in real-world datasets.

V. OPEN RESEARCH CHALLENGES AND RESEARCH IMPLICATIONS

This section addresses the RQ2 and discusses the benefits of SLR with more emphasis towards the open research challenges and research implications towards data intrinsic and class imbalance problem.

In Machine learning, the context of learning imbalanced datasets is considered as a vital issue. From the literature study, it is observed that vast range of research works are concentrated on pre-processing of data before a classifier is generated that in turn provide the best solution than other techniques as it allows dynamic insert or deletes data to balance the dataset.

Besides that, rebalancing of the class distribution manually doesn't have a significant effect on the performance of the classifier.

Class-overlapping is considered to be an alternate cause of class imbalance problem as it yields poor performance of a classifier. In the context of solving class overlapping problem, different sampling techniques are adopted in research studies in the pre-processing phase. These sampling approaches might not be appropriate in solving this class overlapping problem. Several research studies in the literature of the prediction techniques have identified that the class imbalanced problems are caused because of redundant features and the future direction is to address appropriate feature selection techniques to handle class imbalance problems.

Due to the advent of the Internet of Things (IoT) and other networking solutions, a massive amount of data is being generated from various data sources and developing an efficient mechanism for processing such real-world redundant data is considered to be a major challenge in the domain of classification problem. Research works on handling big imbalanced data are focused towards the implementation of machine learning algorithms using Spark and Map Reduce. Handling big real-world datasets that include patient dataset for predicting disease symptoms, financial datasets to predict share markets etc is considered to be open research direction as it includes a large amount of redundant data.

VI. CONCLUSION

The SLR conducted in this article includes a large number of state-of-art research studies addressing the techniques that handle the class imbalance problem intended for the classification in the real world data sets. Several researchers have made significant progress in proposing various techniques to balance the skewed distribution of data. Furthermore, the article addresses the comprehensive taxonomy that classifies the research efforts of the area. A Systematic study of articles from a decade is carefully conducted to analyze the research gaps and propose the research implications for further research in this area. Based on the research results it is observed that this study have exposed interesting research patterns and trends that could be implemented to underline the vital challenges in the field of machine learning to enhance the prediction accuracy.

REFERENCES

1. N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent data analysis 6 (5) (2002) 429–449.
2. L. Rokach, Ensemble-based classifiers, Artificial Intelligence Review 33 (1) (2010) 1–39.
3. R. Anand, K. G. Mehrotra, C. K. Mohan, S. Ranka, An improved algorithm for neural network classification of imbalanced training sets, IEEE Transactions on Neural Networks 4 (6) (1993) 962–969.
4. P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80 (4) (2007) 571 –583
5. X. Xiaoying, F. Sheng, A synthesized sampling approach for improving the prediction of imbalanced classification, in: Proceedings of the IEEE 3rd International Conference on Software Engineering and Service Science (ICSESS), IEEE, 2012, pp. 615–619
6. K. Li, W. Zhang, Q. Lu, X. Fang, An improved smote imbalanced data classification method based on support degree, in: Proceedings of the International Conference on Identification, Information and Knowledge in the Internet of Things, IEEE, 2014, pp. 34–38
7. P. Li, P.-L. Qiao, Y.-C. Liu, A hybrid re-sampling method for svm learning from imbalanced data sets, in: Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Vol.2, IEEE, 2008, pp. 65–69
8. X. T. Dang, D. H. Tran, O. Hirose, K. Satou, Spy: A novel resampling method for improving classification performance in imbalanced data, in: Proceedings of the 7th International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2015, pp. 280–285.
9. F. Koto, Smote-out, smote-cosine, and selected-smote: An enhancement strategy to handle imbalance in data level, in: Proceedings of the 2014.
10. H. F. Oktavino, N. U. Maulidevi, Information extractor for small medium enterprise aggregator, in: Proceedings of the International Conference on Data and Software Engineering (ICODSE), IEEE, 2014, pp. 1–5.
11. Y.-P. Zhang, L.-N. Zhang, Y.-C. Wang, Cluster-based majority undersampling approaches for class imbalance learning, in: Proceedings of the 2nd IEEE International Conference on Information and Financial Engineering (ICIFE), IEEE, 2010, pp. 400–404.
12. J. Wang, Y. Yao, H. Zhou, M. Leng, X. Chen, A new over-sampling technique based on svm for imbalanced diseases data, in: Proceedings of the 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), IEEE, 2013, pp. 1224–1228.
13. Y. Fu, H. Zhang, Y. Bai, W. Sun, An under-sampling method: Based on principal component analysis and comprehensive evaluation model, in: Proceedings of the 2016 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), IEEE, 2016, pp. 414–415.
14. Y. Ma, G. Luo, J. Li, A. Chen, Combating class imbalance problem in semi-supervised defect detection, in: Proceedings of the 2011 International Conference on Computational Problem-Solving (ICCP), IEEE, 2011, pp. 619–622
15. Q. Wu, M. Wang, Abnormal bgp routing dynamics detection by sampling 800 approach in decision tree, in: Proceedings of the 1st International Workshop on Database Technology and Applications, IEEE, 2009, pp. 170–173.

A. Majumder, L. Behera, V. K. Subramanian, Gmr based pain intensity recognition using imbalanced data handling techniques, in: Proceedings of the International Conference on Signal and Information Processing (IConSIP), IEEE, 2016, pp. 1–5.

16. X. Shi, G. Xu, F. Shen, J. Zhao, Solving the data imbalance problem of p300 detection via random under-sampling bagging svms, in: Proceedings of the 2015 International Joint Conference on Neural Networks(IJCNN), IEEE, 2015, pp. 1–5
17. M. Farajzadeh-Zanjani, R. Razavi-Far, M. Saif, Efficient sampling techniques for ensemble learning and diagnosing bearing defects under class imbalanced condition, in: Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2016, pp. 1–7.

A. Ghazikhani, R. Monsefi, H. S. Yazdi, Svbo: Support vector-based oversampling for handling class imbalance in k-nn, in: Proceedings of the 20th Iranian Conference on Electrical Engineering (ICEE), IEEE, 2012,pp.605–610

18. S. Hu, Y. Liang, L. Ma, Y. He, Msmote: improving classification performance when training data is imbalanced, in: Proceedings of the Second International Workshop on Computer Science and Engineering IEEE, 2009, pp. 13–17.
19. R. Liu, R. Huang, Y. Qian, X. Wei, P. Lu, Improving user's quality of experience in imbalanced dataset, in: Proceedings of the 2016 International Conference on Wireless Communications and Mobile Computing Conference (IWCMC), IEEE, 2016, pp. 644–649
20. J. Li, H. Li, J.-L. Yu, Application of random-smote on imbalanced data mining, in: Proceedings of the 2011 4th International Conference on Business Intelligence and Financial Engineering (BIFE), IEEE, 2011, pp.130–133.
21. C. Zhou, B. Liu, S. Wang, Cmo-smote: Misclassification cost minimization oriented synthetic minority oversampling technique for imbalanced learning, in: Proceedings of the 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Vol. 2, IEEE, 2016, pp. 353–358
22. U. Bhowan, M. Johnston, M. Zhang, Differentiating between individual class performance in genetic programming fitness for classification with unbalanced data, in: Proceedings of the IEEE Congress on Evolutionary Computation, IEEE, 2009, pp. 2802–2809.

23. P. Ruangthong, S. Jaiyen, Bank direct marketing analysis of asymmetric information based on machine learning, in: Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, 2015, pp. 93–96.
24. G. Y. Wong, F. H. Leung, S.-H. Ling, An under-sampling method based on fuzzy logic for large imbalanced dataset, in: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE,2014,pp. 1248–1252.
25. T. M. Padmaja, P. R. Krishna, R. S. Bapi, Majority filter-based minority prediction (mfmp): An approach for unbalanced datasets, in: Proceedings of the IEEE Region 10 Conference TENCON 2008, IEEE, 2008, pp. 1–6
26. D. Zhang, J. Ma, J. Yi, X. Niu, X. Xu, An ensemble method for unbalanced sentiment classification, in: Proceedings of the 11th International Conference on Natural Computation (ICNC), IEEE, 2015, pp. 440–445.
27. Y. Fan, Z. Kai, L. Qiang, A revisit to the class imbalance learning with linear support vector machine, in: Proceedings of the 2014 9th International Conference on Computer Science & Education (ICCSE), IEEE, 2014, pp.516–521.
28. S. Chen, H. He, Sera: selectively recursive approach towards nonstationary imbalanced stream data mining, in: Proceedings of the InternationalJoint Conference on Neural Networks, IEEE, 2009, pp. 522–529.
29. P. Sarakit, T. Theeramunkong, C. Haruechaiyasak, Improving emotion classification in imbalanced youtube dataset using smote algorithm, in: Proceedings of the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA),IEEE,2015,pp.1–5.doi:10.1109/ICAICTA.2015.7335373.
30. M. Gao, X. Hong, S. Chen, C. J. Harris, Probability density function estimation based over-sampling for imbalanced two-class problems, in: Proceedings of the 2012 International Joint Conference on Neural Networks(IJCNN),IEEE,2012,pp.1–8. doi:10.1109/IJCNN.2012.6252384.
31. M. Hosseinzadeh, M. Eftekhari, Improving rotation forest performance for imbalanced data classification through fuzzy clustering, in: Proceedings of the 2015 International Symposium on Artificial Intelligence and Signal Processing (AISP), IEEE, 2015, pp. 35–40.
32. H. Shin, S. Cho, How to deal with large dataset, class imbalance and binary output in svm based response model, in: Proceedings of the Korean Data Mining Conference, 2003, pp. 93–107.
33. G. Cohen, M. Hilario, C. Pellegrini, One-class support vector machines with a conformal kernel. a case study in handling class imbalance, in: Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and SyntacticPattern Recognition (SSPR), Springer Berlin, 2004, pp. 850–858.
34. G. Cohen, M. Hilario, H. Sax, S. Hugonnet, Asymmetrical margin approach to surveillance of nosocomial infections using support vector classification, 2003, pp. 1–13
35. G. Wu, E. Y. Chang, Aligning boundary in kernel space for learning imbalanced dataset, in: Proceedings of the 4th IEEE International Conferenceon Data Mining, IEEE, 2004, pp. 265–272.
36. C. Phua, D. Alahakoon, V. Lee, Minority report in fraud detection: Classification of skewed data, SIGKDD Explor. Newsl. 6 (1) (2004) 50–59
37. N. Mustafa, J.-P. Li, Medical data classification scheme based on hybridized smote technique (hst) and rough set technique (rst), in: Pro-ceedings of the IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), IEEE, 2017, pp. 49–55.
38. H. Wei, B. Sun, M. Jing, Balancedboost: A hybrid approach for real-time network traffic classification, in: Proceedings of the 23rd International Conference on Computer Communication and Networks (ICCCN), IEEE,2014, pp. 1–6. doi:10.1109/ICCCN.2014.6911833
39. X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (2) (2009) 539–550
40. C. Ceballes-Serrano, S. Garcia-Lopez, J. Jaramillo-Garz´on, G. Castellanos-Dom´ınguez, A strategy for classifying imbalanced data sets based on particle swarm optimization, in: Proceedings of the 2012 XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), IEEE, 2012, pp. 218–222.
41. D. V. Oliveira, T. N. Porpino, G. D. Cavalcanti, T. I. Ren, A bootstrapbased iterative selection for ensemble generation, in: Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), IEEE, 2015, pp. 1–7.
42. P. Ruangthong, S. Jaiyen, Hybrid ensembles of decision trees and Bayesian network for class imbalance problem, in: Proceedings of the 8th International Conference on Knowledge and Smart Technology (KST), IEEE, 2016, pp. 39–42.
43. Q. Liu, X. Wei, R. Huang, H. Meng, Y. Qian, Improved adaboost model for user's qoe in imbalanced dataset, in: Proceedings of the 2016 8th International Conference on Wireless Communications & Signal Processing (WCSP), IEEE, 2016, pp. 1–5.
44. Q. Huang, X. Zhang, An improved ensemble learning method with smote for protein interaction hot spots prediction, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016, pp. 1584–1589.
45. B. Pal, M. K. Paul, A gaussian mixture based boosted classification 1105 scheme for imbalanced and oversampled data, in: Proceedings of the International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2017, pp. 401–405.
46. X.-S. Hu, R.-J. Zhang, Clustering-based subset ensemble learning method for imbalanced data, in: Proceedings of the 2013 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 1, IEEE, 2013, pp.35–39
47. G. Mustafa, Z. Niu, A. Yousif, J. Tarus, Distribution based ensemble for class imbalance learning, in: Proceedings of the Fifth International Conference on Innovative Computing Technology (INTECH), IEEE, 2015, pp. 5–10
48. G. G. Sundarkumar, V. Ravi, V. Siddeshwar, One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection, in: Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), IEEE, 2015, pp. 1–7.
49. B. Krawczyk, G. Schaefer, An analysis of properties of malignant cases for imbalanced breast thermogram feature classification, in: Proceedings of the 2013 2nd IAPR Asian Conference on Pattern Recognition(ACPR), IEEE, 2013, pp. 305–309.
50. Z. Yongqing, Z. Min, Z. Danling, M. Gang, M. Daichuan, Improved smotebagging and its application in imbalanced data classification, in: Proceedings of the IEEE Conference Anthology, IEEE, 2013, pp. 1–5.
51. X. Wu, S. Meng, E-commerce customer churn prediction based on improved smote and adaboost, in: Proceedings of the 13th International Conference on Service Systems and Service Management(ICSSSM),IEEE, 2016, pp. 1–5.
52. N. Thai-Nghe, A. Busche, L. Schmidt-Thieme, Improving academic performance prediction by dealing with class imbalance, in: Proceedings ofthe 9th International Conference on Intelligent Systems Design and Applications, IEEE, 2009, pp. 878–883.
53. K. Gao, T. Khoshgoftaar, A. Napolitano, Improving software quality estimation by combining boosting and feature selection, in: Proceedings of the 12th International Conference on Machine Learning and Applications(ICMLA), Vol. 1, IEEE, 2013, pp. 27–33.
54. S. M. El-Ghamrawy, A knowledge management framework for imbalanceddata using frequent pattern mining based on bloom filter, in: Proceedingsof the 2016 11th International Conference on Computer Engineering &Systems (ICCES), IEEE, 2016, pp. 226–231.
55. S. Soltani, J. Sadri, H. A. Torshizi, Feature selection and ensemble hierarchical cluster-based under-sampling approach for extremely imbalanced datasets: Application to gene classification, in: Proceedings of the 1st International eConference on Computer and Knowledge Engineering (ICCKE), IEEE, 2011, pp. 166–171.
56. Zughrat, M. Mahfouf, S. Thornton, Performance evaluation of svm and iterative fsvm classifiers with bootstrapping-based over-sampling and under-sampling, in: Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2015, pp. 1–7.
57. T. Sandhan, J. Y. Choi, Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition, in: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR), IEEE, 2014, pp. 1449–1453.
58. G. I. Winata, M. L. Khodra, Handling imbalanced dataset in multi-label text categorization using bagging and adaptive boosting, in: Proceedingsof the 2015 International Conference on Electrical Engineering and Informatics, IEEE, 2015, pp. 500–505.
59. E. Dwiyanti, Adiwijaya, A. Ardiyanti, Handling imbalanced data in churn prediction using rusboost and feature selection (case study: Pt.telekomunikasiindonesia regional 7), in: Proceedings of the 2nd International Conference on Soft Computing and Data Mining, Springer, 2017, pp. 376–385.

60. M. Galar, A. Fern´andez, E. Barrenechea, F. Herrera, Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling, Pattern Recognition 46 (12) (2013) 3460–3471.
61. Fern´andez, S. del R´ıo, N. V. Chawla, F. Herrera, An insight into imbalanced big data classification: outcomes and challenges, Complex & Intelligent Systems (2017) 1–16
62. Y. Qian, Y. Liang, M. Li, G. Feng, X. Shi, A resampling ensemble algorithm for classification of imbalance problems, Neurocomputing 143 (2014) 57–67.
63. Q. D. Tran, P. Liatsis, Raboc: An approach to handle class imbalance in multimodal biometric authentication, Neurocomputing 188 (2016) 167–177.
64. R. Barandela, R. M. Valdovinos, J. S. S´anchez, New applications of ensembles of classifiers, Pattern Analysis & Applications 6 (3) (2003) 245–256.
65. H. Guo, H. L. Viktor, Learning from imbalanced data sets with boosting and data generation: the databoost-im approach, ACM Sigkdd Explorations Newsletter 6 (1) (2004) 30–39.
66. K. Jiang, J. Lu, K. Xia, A novel algorithm for imbalance data classification based on genetic algorithm improved smote, Arabian Journal for Science and Engineering 41 (8) (2016) 3255–3266.

AUTHORS PROFILE

K. Santhi, received her B.Tech degree from Vignan's Engineering College, JNTUH, Guntur in 2003, M.Tech in computer Science and Engineering from JNTU, Hyderabad in 2006 and pursuing Ph.D. in Computer Science and Engineering from S V University College of Engineering, Tirupati in 2016. Her areas of interest are Big Data Analytics, Machine Learning , Data Mining.

Dr. A. Rama Mohan, Reddy was born in1958, received his B.Tech degree from JNT University Anantapur in 1986, Masters in Computer Science and Engineering from NIT, Warangal in 1991 and Ph.D. in Computer Science and Engineering from Sri Venkateswara University, Tirupati in 2007.He is currently working as a Professor of Computer Science and Engineering, S V University College of Engineering, Tirupati, India. His research interests are Software Engineering, Software Architecture, Cloud Computing, Operating Systems and Data Mining.