An Integrated Framework for Knowledge Extraction using Clustering and Classification

ABSTRACT

Data is increasing at an unimaginable rate every year. The area of Data Mining has arisen over the last decade to address this problem. Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas starting from simple applications like supermarket transactions, railway reservations to the more complex and complicated ones like space research, molecular databases, images, and astronomical bodies etc. Using this data to discover hidden knowledge, unexpected patterns and unknown information is Data Mining.

Data Mining includes two indispensable tasks, Clustering, and Classification. The integration of these tasks together can undertake challenging machine-learning problems. Taking advantage of these methods is a significant research area. There are few quality issues present in most of the real-world datasets that negatively influence the performances of the Classifier, such as noisy and incomplete Data, Outliers, High Dimensional Data, and Class imbalanced Distribution (Between-Class imbalanced and Within-Class imbalanced). The research carried out here provides a generic and logical solution to the above-mentioned problems and has been dealt with separately in subsequent chapters. The models are presented, evaluated, and compared for each of the above-mentioned issues on various performance measuring parameters with State-of-the-art methods. The first proposed model provides a Cluster-based approach using an under-sampling solution to balance the imbalanced data. A study for binary class distribution on 12 data sets openly available in UCI machine learning repository with different degrees of imbalance nature has been conducted that presented a three-fold solution. First, it balances imbalanced data using the K- means clustering approach. The main purpose of this approach is to selectively discard majority instances from the dataset to make the distribution balanced, which can be applied to any traditional classifier. Secondly, it is capable to handle between class imbalance distribution and within-class distribution. Thirdly, it can handle the different degrees of imbalanced distribution. The proposed method is simple yet effective in order to classify the imbalanced distribution of Data.

The second proposed model is to improve the performance of any classifier using a hybrid model (prior clustering to classification). The research experiments observed that the hybrid model is more refined and accurate than a single individual classifier. The objective is to utilize the strength of one method to complement the weaknesses of another. If the interest is to find the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model. The experiment has been conducted on 13-benchmark datasets taken from the UCI machine learning repository. The results evidently revealed that this integration process generates a more precise and accurate model.

The third proposed model is applicable as a Feature compression and Extraction technique to build a better feature space because it can solve a number of machine learning problems. High dimensional data generally diminish the accuracy and efficiency of Data Mining algorithms. The advantage of the proposed model is; it first identifies the relevant features that may lead to accurate results and then classifies it. The experiments were conducted on nine-benchmark dataset taken from the UCI machine repository with diverse degrees of dimensionality. The Comparative Analysis of the Proposed Approach with a standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) was presented on these datasets.

The developed models have been published in Conference Proceedings / International Journals and the details of the publications are mentioned in the end.

An Integrated Framework for Knowledge Extraction using Clustering and Classification

TABLE OF CONTENTS

1 Introduction 5

1.1 Overview of Data Mining 5

1.2 Data Mining Applications summarized 6

1.3 The Data Mining Process 6

1.4 Data mining task 7

1.4.1 Predictive mining task 7

1.4.2 Descriptive task 8

1.5 Problem Formulation 8

1.6 Research Objectives 9

1.7 Research contribution 9

2 Literature Review 11

2.1 Literature Review on Machine Learning Techniques 11

2.2 Literature Review on Data Mining Tools & Techniques 12

2.3 Data mining applications in Healthcare 12

2.4 Data mining using clustering and classification Integrated approach 13

3 A Cluster-based solution for imbalanced Data 15

3.1 Introduction 15

3.1.1 Methods of handling imbalanced Data 15

3.2 Conclusion 23

4 A Hybrid Model to Enhance the Performance of a Classifier 24

4.1 Introduction 24

4.2 Datasets Description 24

4.3 Methodology 24

4.3.1 Data Preparation 24

4.3.2 Clusters Building 25

4.3.3 Building the classification Models: 25

4.4 Algorithm 27

4.5 Results & Conclusion 27

5 Clustering-based Feature Selection method 29

5.1 Introduction 29

Introduction about Data Set 29

5.2 Methodology 30

5.3 SVM Classifier: 30

5.4 Performance analysis of the Generated Model 32

5.5 Results & Conclusion 32

5.6 Further Enhancements 33

6 References 34

7 Publications 38

List of Figures

Sr. No.	Figure Number	Figure Caption	Page No.
1	1.1	The KDD Process	7
2	1.2	Data Mining Task	7
5	3.1	Random Under-Sampling & Random Oversampling	15
6	3.2	The framework of Cluster-based Sampling	17
7	3.3	Flow chart of the proposed algorithm	20
8	3.4	Charts for Accuracy	21
9	3.5	Charts for F-measure	22
10	3.6	Chart for F-P rate	22
11	3.7	Chart for Precision	22
12	3.8	Chart for ROC	23
13	4.1	Framework for the hybrid model	25
14	4.2	Charts for the Accuracy of classifiers on five values of K for different datasets	28
16	5.1	Flow chart for the proposed model	30
17	5.2	Chart for Heart disease dataset	32
18	5.3	Chart for Colon cancer dataset	32
19	5.4	Chart for Image segmentation	32
20	5.5	Chart for Cleland dataset	32

List of Tables

Sr. No.	Table Number	Table Caption	Page No.
1	Table 1.1	Data Mining Application areas	6
2	Table 3.1	Imbalanced Datasets with different degrees of imbalanced Distribution	16
3	Table 3.2	Overall performance of the model on various parameters	21
4	Table 4.1	Data set Description	24
5	Table 4.2	Proposed model performance on different cluster numbers	26
6	Table 5.1	Data set description	29
7	Table 5.2	Comparative performance analysisof proposed approach	31

Chapter-1

1 INTRODUCTION

1.1 OVERVIEW OF DATA MINING

"If you mine the data hard enough, you can also find messages from God"
— Scott Adams, Dilbert

The world today is flooded with data; since human life began, hunters seek patterns in animal's migration behavior. Farmers seek patterns in crop growth, Politicians seek patterns in voter opinion's, Users create content such as Blog posts, tweets, social-network interactions, and photographs; servers continuously create activity logs, the internet becomes the repository of data. Over the last few decades, data growth has been on a massive scale that has become problematic for scalability. This brought about a challenging situation for the handling of this huge some of the data.

Exponentially increasing volumes of data with lots of complexity require effective and automatic approaches to inference knowledge from this voluminous data are important. Extraction methods with high efficiency are needed to gain valuable knowledge from the hidden or undiscovered patterns amongst colossal amounts of data. Efficient means of storing, retrieving and manipulating data, as a revolution in information availability and exchange via the internet helps the database technologists to focus on developing techniques for learning and acquiring knowledge from the data. The efficient decision-making process is the key to a successful organization that is based on timely and valuable knowledge.

It is important to understand the difference between a model and a pattern that helps in understanding the structure, relationships to a relatively small part of the data or the space in which the data would occur "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [1] is the definition of Data Mining. It is also a science of extracting useful information from large data sets or databases.

The definition of Data Mining as per Hand et al is "a well-defined procedure that takes data as input and produces output in the forms of models or patterns". Data mining is being used in many fields such as Healthcare system, Biological systems, Medical sciences, Biometric applications, Pharmaceutical industries, banking, marketing analysis, Multimedia system, Retail and e-commerce, Spatial data analysis, Security system, String mining, fraud and intrusion detection, image processing, telecommunication industry, scientific applications, etc.[1, 2].

Raw data is extracted through data mining techniques, which is useful in the process of prediction analysis, clustering that helps in the generation of many data mining techniques & tools [3]. When the abundance of data is available, it can be the past & future data, which gets analyzed by data analysis. Data mining is a multi-disciplinary field includes the combination of statistics, machine learning, artificial intelligence and database technology and applications of it is very high[4].

1.2 DATA MINING APPLICATIONS SUMMARIZED

Table 1.1 Data Mining application areas

Applications	Usage
Communications	In communication for the prediction of customer behavior for the relevant campaigns and target marketing.
Insurance	Helps improve their profits through the implementation of offers to the existing & new customers.
Education	In this sector, it helps in obtaining data of students, their profile and weak students in a particular subject can be sorted out easily.
Manufacturing	Manufacturers get the predictive approach towards the working & maintenance of machineries that helps them prevent major damages.
Banking	Data mining is useful to banks in the identification of defaulters, loans, etc and helps them to keep track of the banking details.
Retail	In malls & retail sectors of groceries, it helps in the sorting out of the highest selling commodity that attracts customer's attention and thus increases its profits.
Service Providers	Prediction of customer behavior through their billing details, interaction at service centers & complaints will improve the approach of mobile & utility industries in offering incentives to the customers.
E-Commerce	E-commerce websites use Data Mining for the promotion of their sales so that more customers are attracted towards e-commerce. This improves their sales strategy.
Super Markets	It allows supermarkets to predict the purchasing patterns of their shoppers & target or offer specific products according to their needs.
Crime Investigation	In crime investigation, it helps the police force to get deployed based on the likeliness of crime in a particular area.
Bioinformatics	In this field, it helps in the segregation of specific data from the biological & medicinal data sets.
Health care	Data mining is highly helpful in health care to help in decision making about treatment, healthcare, Customer Relationship. Apart from this, it is helpful in various related sectors of pharmaceutical industries, medical device industries, etc.[5]

1.3 THE DATA MINING PROCESS

As we have seen, the amount of data in databases is increasing at a tremendous rate. This growing need gives birth to a new research field to aid humans too intelligently and, automatically analyze huge data sets called Knowledge Discovery in Databases (KDD). Researchers start giving attention to many different fields including pattern recognition, database design, machine learning, statistics, and data visualization [6]. The KDD process can be decomposed into the following steps as illustrated in Figure 1.1[7]:

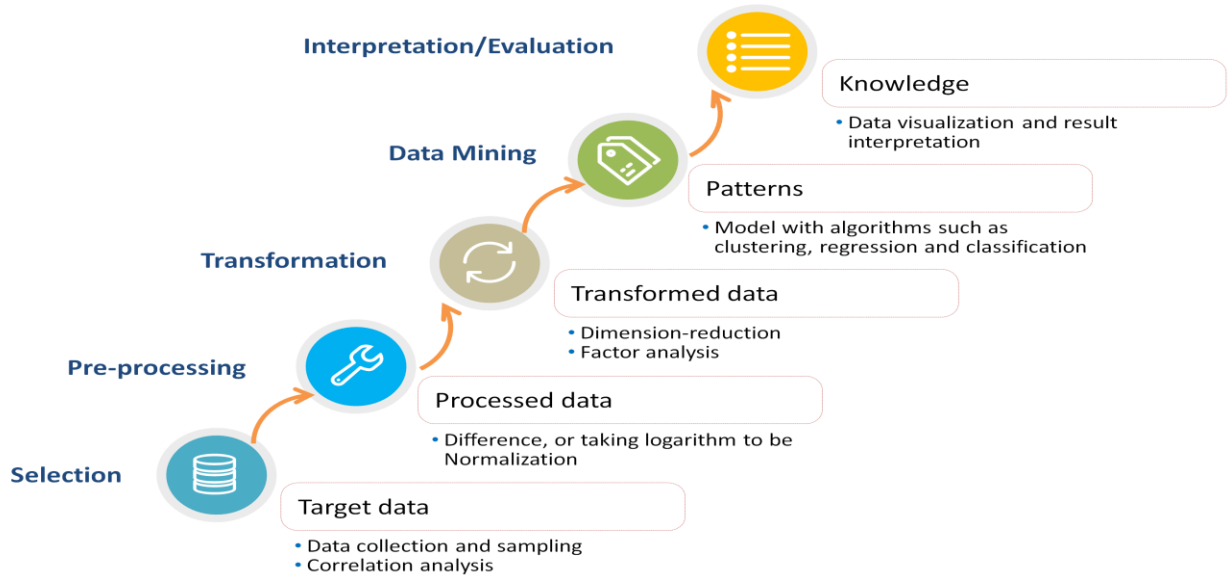


Figure 1.1: KDD Process **source: Fayyad (1996)**

1.4 DATA MINING TASK

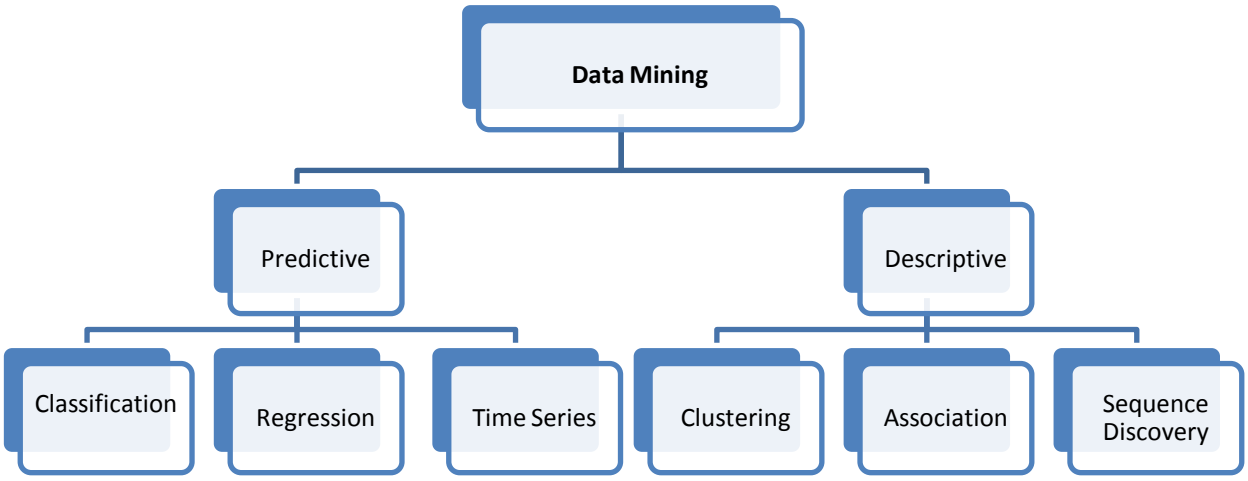


Figure 1.2: Data Mining Task

In general, the Data-Mining task is of two types Predictive & Descriptive.

1.4.1 PREDICTIVE MINING TASK

It is a task wherein we obtain inference on the present data to make predictions. The most popular predictive Data Mining task is Classification. The process of finding a model depends on the analysis of training data whose class label is known. The purpose of this model is to predict the unknown class label from the class of objects. The target dataset is randomly divided into two mutually exclusive and exhaustive sets called training set and test set to carry out classification. The relationship between the conditional attributes and the decision attribute is derived from the training data, by which a model (or a function) is derived to describe the concepts. Such a model can be represented in various types such as mathematical formulae, decision trees, prediction (if … then) rules, or neural networks.

For the prediction of the class of each data instance, this model is used in the test set. Supervised data mining is Prediction. Explicitly or implicitly constructed model is by generalized by a sufficient number of training examples, which are based on inductive learning of most data mining techniques. This is known as a generalization of knowledge and

the trained model is assumed to be applicable in future cases. Decisions can be made with the help of predictions made. Knowledge discovery in databases predicts unknown or future values of some attributes based on other attribute values in the database through predictive modeling. [8] Predefined classes of any instance could be predicted through classification in data mining techniques.

- **Classification**

Classification is a data mining technique used to predict the class to which an instance belongs to predefined classes. Classification is supervised learning. It tries to assign previously unseen records to a class as accurately as possible. Classification is a two-stage method where at the first stage we built/trained models from historical training data sets with labeled class attributes. In the second stage, we try to maximize the classification accuracy rate which is the ratio of the number of correct predictions to the total number of predictions in the test dataset. Fundamentally, it is a mapping from target function to attribute set of already labeled class [9].

Classification provides a model that can describe future data. In a collection of training samples, the type of task could be designed for a model with class attributes as a function with values of other attributes (Duda et al., 2001)[10]An existing unseen record is segregated to an accurate class. The classification task is to maximize the accuracy rate that is the ratio of the number of correct predictions to the total number of predictions in the test dataset. [11]

1.4.2 DESCRIPTIVE TASK

The objective of this task is to derive patterns to summarize the underlying relationship and amongst data.

- **Clustering**:

The most efficient technique that is applied to raw data for the extraction of useful information Similar and dissimilar types of data are clustered for analysis of useful information from the dataset. When data is sorted into groups of similar objects it is known as Clustering. Each one of such groups is known as a cluster. Every cluster has similar objects but differs from the objects of the other clusters (Hoppner et al., 1988) [13]. Clustering is not predefined but an unsupervised learning method & class labels are not present.

Clustering similar to classification in which objects of data are grouped without consulting a known class label. Data groupings are not pre-defined in clustering; they are generated by the similarities within the data objects based on the characteristics present in the actual data. Partition or division of datasets into clusters or many groups is based on their similarities that are done in such a way that the objects with maximum similarities belong to one group or cluster and are highly dissimilar with the other group or cluster. That means a good clustering algorithm has a maximum intra-cluster similarity and minimum inter-cluster similarity. Data clustering does not require category labels or predefined group information, unlike data classification. Clustering is studied in the field of machine learning as an unsupervised learning process, as it is "learning from observation" and not "learning from examples." The pattern proximity matrix is being measured as a distance function defined on pairs of patterns (Jain & Dubes, 1988; Duda et al., 2001) [14]

1.5 PROBLEM FORMULATION

Clustering gives an overview of a given data set, insight into the data distribution within a data set is often sufficient. Preprocessing for other data mining algorithms is another important use of clustering algorithms. Clustering does not require a specification of a set of examples, which is its special feature. Thus, clustering is applied in applications in which no or little prior knowledge of the groups or classes in a database is available. The clustering usefulness is often associated with individual interpretation and on the selection of suitable similarity measure. [15]

Classification is necessary to segregate data into predefined classes. It depends on the attributes and features present in the data. Users can get a better understanding and description of the data for each class of the database.

There are few quality issues, that negatively influence the performances of the classifier, such as Noisy and incomplete data, high dimensional data, the performance of a classifier, and class imbalanced distribution (Between-class imbalanced and within-class imbalanced).

A vast amount of data is not possible to process or analyze by one classifier. Hence, it can be partitioned into smaller chunks and these smaller chunks are passed to multiple classifiers ensemble, rather than one classifier. The unseen data can be passes to different classifiers for classification once the learning over a different chunk of data is done or the model is ready.

1.6 RESEARCH OBJECTIVES

The overall objective of this study is to develop the most accurate and understandable model by combining supervised (Classification) and unsupervised (Clustering) learning methods. This study has proved that the integrated model does not only improve the accuracy of the classifier but it can handle some challenging problems exist in Data mining areas for years. The integration of these methods can serve the following purposes.

Feature compression & extraction:

- Based on clustering criteria features can be clustered together. It reduces dimensionality, complexity and computation time, and increases comprehensibility and the overall performance of classification algorithms.

A fully labeled training set from the unlabeled set can be created:

- To get a fully labeled data is difficult when we label it manually by human expertise; it is expensive, time-consuming and error-prone. In such cases, Clustering is used to label unlabeled information to boost the performance of the classification task.

Improve the performance of the classifiers:

- The objective is to utilize the strength of one method to complement the weaknesses of another. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as an integrated model.

Imbalanced Data to balance:

- imbalanced class distribution is a common problem in real-life applications. Learning from imbalanced data is identified as an open research problem for decades. It affects the performance of standard classifiers so drastically due to the unequal distribution of data among classes.

1.7 RESEARCH CONTRIBUTION

Three models are presented and implemented those offer a generic and logical solution to meet up the research objectives.

The first proposed model provides a Cluster-based approach using an under-sampling solution to balance the imbalanced data. A study for binary class distribution on 12 data sets Abalone, Cleveland-0, Ecoli-3, Glass-1, Haberman, New-Thyroid-1, Page-Blocks-0, Pima, Wine Quality White, Breast Cancer, Wisconsin, Yeast 1, and Vowel openly available in UCI machine learning repository with different degrees of imbalance nature is conducted. The proposed framework is capable to give a three-fold solution. First, it balances imbalanced data using the K-means clustering approach. The main purpose of this approach is to selectively discard majority instances from the dataset to make the distribution balanced and can be applied to any traditional classifier. Secondly, it is capable to handle Between-class imbalance distribution and within-class distribution. Thirdly it can handle the different degrees of imbalanced distribution. The proposed method is simple yet effective in order to classify the imbalanced distribution of Data.

The second proposed model is to improve the performance of any classifier using a hybrid model (prior clustering to classification). The research experiments observed that the hybrid model is more refined and accurate than a single individual classifier. The objective is to utilize the strength of one method to complement the weaknesses of another. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model. The proposed method consists of three steps. Clustering Step: Prior to the classification, the dataset is grouped into the different number of clusters using the K-means algorithm depending upon the value of K (K=2, 3, 4 and 5). Expansion step: The resultant dataset augmented with one more feature originated from the clustering step named cluster_no as an input to the classifier. Classification step: The expanded data is applied to SVM (Support Vector Machine) classifier. The experiment conducted on 13 benchmark dataset taken from UCI machine learning repository, these are- Heart Disease dataset, Wine quality white dataset, Parkinson dataset, Colon Cancer dataset, Breast Cancer dataset, Image Segmentation dataset, Cleveland-0 Dataset, Ionosphere Dataset, Squash Harvest Stored dataset, Bank Dataset, Glass Dataset, Pima dataset, and Haberman Dataset. The results revealed that this integration process generates a more precise and accurate model. All classifier's accuracy gets affected positively when supervised and unsupervised learning methods are combined.

The third proposed model is applicable as a Feature compression and Extraction technique to build a better feature space because it can solve a number of machine learning problems. High dimensional data generally diminish the accuracy and efficiency of Data Mining algorithms. The higher the dimensionality, the higher the computation cost involved in processing. Irrelevant features may exert undue burden on classification evaluation parameters and increases the time and resources needed to build the model. The advantage of the proposed model is, it first identifies the relevant features that may lead to accurate results. The experiments were conducted on nine benchmark dataset taken from UCI machine repository with diverse degrees of dimensionality they are Heart Disease, Wine quality white, Parkinson dataset, Colon Cancer, Breast Cancer, Image Segmentation, Cleveland-0, Ionosphere and Squash Harvest Stored. The Comparative Analysis of the Proposed Approach with a standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) was presented on the aforementioned datasets. The analysis was presented on 20, 40, 60, 80 and 100 % of the features on three performance measuring parameters; Accuracy, F-Measure and ROC. The K-means algorithm is used to group similar features together. The centroid of each cluster has been taken as the individual cluster representative. It reduces dimensionality, complexity and computation time, and increases comprehensibility and the overall performance of classification algorithms. Therefore, the proposed model first identifies the relevant features that may lead to accurate results.

CHAPTER-2

2 LITERATURE REVIEW

There are Rapid technological inventions in the Data mining domain at an extraordinary pace. Its implementation of real-world problems in diverse areas – arises a problem in several aspects.

2.1 LITERATURE REVIEW ON MACHINE LEARNING TECHNIQUES

Mduma et al. [16] reported the highlights of open challenges for future research directions. Therefore, in this article, a survey of how machine-learning techniques have been used in the fight against dropouts is presented. The purpose of the conducted survey is to provide a stepping-stone for students, researchers, and developers who aspire to apply the techniques.

Chen, F [17] presented a systematic way to review data mining in knowledge view, technique view and application view, including classification, clustering, association analysis, time series analysis, outlier analysis, etc. In addition, the latest application cases are surveyed. At last, a suggested big data mining system is proposed.

The main purpose of the study is to introduce the basic and core idea of each commonly used clustering algorithm, specify the source of each one, and analyze the advantages and disadvantages of each one were reported in Ann. Data. Sci [18].

AmatulZehra et al.[19] they have focused on the importance of data preprocessing for data mining .the results show a significant improvement in the accuracy of preprocessed data over not preprocessed data.

Anil K. Jain[20], provided a brief overview of clustering, summarize well-known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature selection during data clustering, and large scale data clustering.

Mariscal, G., et al [21] described the most used (in industrial and academic projects) and cited (in scientific literature) data mining and knowledge discovery methodologies and process models, providing an overview of its evolution along data mining and knowledge discovery history and setting down the state of the art in this topic.

S.B. Kotsiantis [22] studied various supervised machine learning classification techniques, of course, a single article cannot be a complete review of all supervised machine learning classification algorithms (also known induction classification algorithms), yet we hope that the references cited will cover the major theoretical issues, guiding the researcher in interesting research directions and suggesting possible bias combinations that have yet to be explored.

RuiXu, [23] surveyed clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. Several tightly related topics, proximity measures, and cluster validation are also discussed.

Sherry Y [24] studied the impacts of data mining by reviewing existing applications, including personalized environments, electronic commerce, and search engines. For these three types of applications, how data mining can enhance their functions is discussed. The reader of this paper is expected to get an overview of the state of the art research associated with these applications. Furthermore, we identify the limitations of current works and raise several directions for future research.

Jun Lee et al. [25] discusses the requirements and challenges of DM and describes major DM techniques such as statistics, artificial intelligence, decision tree approach, genetic algorithm, and visualization.

Dietterich T.G. [26] reported about the original ensemble method is Bayesian averaging, but more recent algorithms include error-correcting output coding, Bagging, and boosting. This paper reviews these methods and explains why ensembles can often perform better than any single classifier.

2.2 LITERATURE REVIEW ON DATA MINING TOOLS & TECHNIQUES

Begum Çıgsar[27] Studied to identify data mining classification algorithms and use them to predict default risks, avoid possible payment difficulties, and reduce potential problems in extending credit. The data for this study, which contains demographic and socioeconomic characteristics of individuals, were obtained from the Turkish Statistical Institute 2015 survey. Six classification algorithms—Naive Bayes, Bayesian networks, J48, random forest, multilayer perceptron, and logistic regression—were applied to the dataset using WEKA 3.9 data mining software.

A literature survey on the important role played by data mining tools in the analysis of the huge volume of healthcare-related data was carried out by Divya Sharma et al.,[28] in the prediction and diagnosis of lifestyle diseases. The objective of their work is to provide a study of different data mining techniques that can be employed in automated lifestyle disease prediction systems.

Amit Verma[29] described the main aim of this paper is to improve information retrieval activities to a higher level. Different methods for information retrieval have been studied and discussed. It involves the use of the Fuzzy Ontology Generation framework (FOGA) framework along with Formal Concept Analysis (FCA) based clustering and keyword matching approach.

Sethunya R Joseph, et al.,[30] researched on data mining has successfully yielded numerous tools, algorithms, methods, and approaches for handling large amounts of data for various purposeful uses and problem-solving.

"A Comparison Study between Data Mining Tools over some Classification Methods" by Abdullah H. Wahbeh et al. [31] helped me to gain knowledge about the implementation of classification algorithms in these tools and gave me a fair idea on how I should work on my study. This research was very useful for me to build a platform for my comparison study.

Rathee A. and Mathur R. P. [32] in this paper, the study of education database is done, which contains hidden knowledge for improving student's performance. This paper gives the comparative study of decision tree algorithms like ID3, C4.5, and CART and as a result, C4.5 is more accurate. The predictions obtained from the algorithms help the teacher to identify poor students and improve their performance.

Rohit Arora [33], used two classification algorithms J48 (which is a java implementation of C4.5 algorithm) and multilayer perceptron alias MLP (which is a modification of the standard linear perceptron) in the Weka interface. It can be used for testing several datasets.

Xindong Wu et al [34] KNN classification is an easy to understand and easy to implement classification technique.

M. A. Hearst et al.[35], in an introductory overview, points out that a particular advantage of SVMs over other learning algorithms is that it can be analyzed theoretically using concepts from computational learning theory, and at the same time can achieve good performance when applied to real problems

2.3 DATA MINING APPLICATIONS IN HEALTHCARE

Hlaudi daniel masethe et al[36], and many others have studied the Heart disease diagnosis that has been done by various data mining methods. Heart disease diagnosis using classification methods such as J48, REPTREE, Naïve Bayes, Bayes Net, Simple CART.Hlaudi Daniel Masethe.Heart Attack Prediction System using the K- means clustering algorithm on the pre-processed data and the recurrent patterns are mined with the MAFIA algorithm Shantakumar B. Patil applied [37].

Jayaram et al.[38] develop a hybrid model for classifying Pima Indian Diabetic Database (PIDD). The model consisted of two stages. Experimental results signify that cascaded K-means clustering and the rules generated by cascaded C4.5 tree with categorical data are easy to interpret as compared to rules generated with C4.5 alone with continuous data.
Type II Diabetes Prediction using Clustering Vijayalakshmi et al. [39] developed a clustering algorithm that is used for predicting diabetes based on graph b-coloring technique. They implement and perform experiments by comparing their approach with K-NN classification and K- means clustering. The results showed that the clustering based on graph coloring is much better than other clustering approaches in terms of accuracy and purity. The proposed technique presented a real representation of clusters by dominant objects that assures the inter-cluster disparity in partitioning and used to evaluate the quality of clusters.
Jason D. M. Rennie et al.,[41], In this paper they propose simple, heuristic solutions to some of the problems with Naive Bayes classifiers, addressing both systemic issues as well as problems that arise because the text is not actually generated according to a multinomial model.

2.4 DATA MINING USING CLUSTERING AND CLASSIFICATION INTEGRATED APPROACH

Hua-Jun Zeng et al [43] presented a clustering-based classification (CBC) approach when sufficient labeled data is not present as most of the traditional text classification algorithm's accuracy degrades drastically on an insufficient amount of labeled data. They have proposed an approach that is more effective with small training data and at the same time easier to achieve high performance especially when the labeled data is sparse. They have cascaded K-means and TSVM clustering and classification algorithms and applied them on 20-Newsgroups, Reuters-21578 and Open Directory Project (OPD) WebPages. The approach initially makes clusters of labeled and unlabeled data and according to the cluster, results expand the labeled set because of more the labeled data greater the accuracy. They have used P=100% for his experiments (means all unlabeled data are labeled by the clustering results). The result reveals that the performance of the CBC is superior over the existing method when labeled data size is very small in all other cases the algorithm is not adaptable.

Asha T, S [44] This paper proposes a combined approach of clustering and classification for the detection of tuberculosis (TB) patients'-means clustering algorithm with Naïve-Bayes, C4.5 decision tree, Support Vector Machine, Ada-Boost, and Random Forest tree are combined to improve the accuracy. Initially, the data is clustered in two sets and respective classes are assigned to them. Then various classification algorithms are trained with these clusters based on the K fold cross-validation method. The best-obtained accuracy is 98.7% with Support Vector Machine is reported.

Shekhar R [45] they have proposed a cascaded method for classifying anomalous activities in a computer network. Two rules are deployed to combine K- means and ID3(The Nearest Neighbor rule 2) the nearest consensus rule.
Mohammad Salim [46], the paper presents an integrated approach of subspace clustering and K nearest neighbor for text classification. The authors proposed an innovation by applying the impurity component for measuring dispersions and chi-square statistics for dimensions of the cluster. The experiments are conducted on NSF abstract datasets and 20 Newsgroup datasets .the comparisons are done on ROC (Receiver Operating Characteristics curve). The proposed method achieved an AUC (Area under the ROC curve) value of .927 while other methods recorded highest as .89 for NSF abstract datasets and for 20 newsgroup datasets AUC value is .813 while other can achieve .77 as highest.

M.I. López [47] Paper presented an approach to predict the performance of the students based on the usages of Forum data. The objective of their study is to determine whether participation in the courses forum can be a good predictor to evaluate the performance in the

examination and how well the proposed integrated model performs over the traditional classification approach on the forum data usages. The outcomes show that student participation in the course forum is a good predictor of the result of the course and the proposed integrated approach. The highest results are obtained by naïve bays with six attributes is 89.4%.

Antonia Kyriakopoulou [48] addressed the incorporation of clustering as a complementary step to text classification and the featured representative of the texts boost the performance of SVM/TSVM classifier experiments that were carried out on ECML/PKDD discovery challenges 2008 for SPAM detection in Social bookmarking system.
B.V. SUMANA [49]They have proposed a hybrid model by cascading clustering and classification .here they have used K- means with a 10 fold cross-validation preprocessing algorithm with 12 classifiers on 5 different medical datasets. They have used a best-first search (BSF) and correlation-based feature selection (CFS) for relevant feature selection.

Chapter-3

3 A CLUSTER-BASED SOLUTION FOR IMBALANCED DATA

3.1 INTRODUCTION

Two indispensable tasks of data mining are clustering & classification[50-52]. The integration of these tasks together can give better and accurate results compared to-unaccompanied. Taking advantage of these methods is a significant research area. There are few quality issues, that negatively influence the performances of the classifier, such as Noisy and incomplete data, outliers, high dimensional data, and class imbalanced distribution (Between-class imbalanced and within-class imbalanced).learning from imbalanced data is one from top 10 challenging problems in data mining[53]. Most of the work is done on - between class imbalance problems. Very few researchers have addressed the problem of imbalanced distribution among data - within the class. Our study is an approach to deal with these two problems (between- class imbalance distribution of data & within-classes imbalance distributions of data) simultaneously.[54-60] This study proposed a cluster-based sampling solution to classify the imbalanced data. K-means algorithm[61 & 62] is used with SVM classifier[63] and observed results proved that the proposed method is simple yet effective in order to classify the imbalanced distribution of data.

3.1.1 METHODS OF HANDLING IMBALANCED DATA

Two methods provided in the literature to tackle imbalanced class distribution.

3.1.1.1 DATA LEVEL APPROACH

In this solution, data is modified to be applied to traditional classifiers**.**

- **Random sampling Techniques**: The most common sampling methods are: random oversampling and random under-sampling. Random oversampling increases the minority class instances, by randomly reproducing the minority class instances. While, Random under sampling reduces - the majority class by randomly removing some majority class instances. Over-sampling increases training time and over-fitting .under-sampling works better compare to over-sampling in terms of both time and memory complexity [64]. Fig. 3 depicts how instances are randomly selected to increase or decrease the sample size.

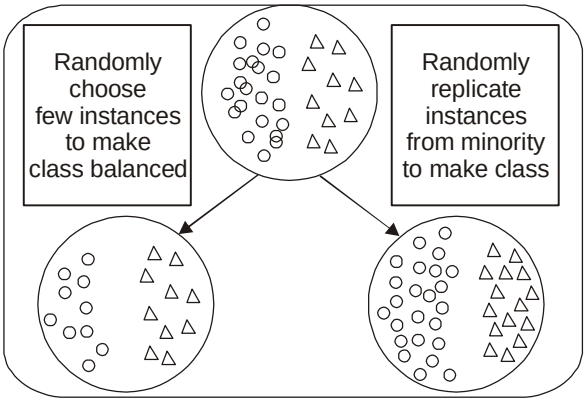


Figure 3.1 Random Under-Sampling and Random Oversampling

- **Synthetic Minority Oversampling Technique** (SMOTE): Chawla *et al*., (2002) [65] proposed a Synthetic Minority Oversampling Technique (SMOTE) - remarkable research in the area of oversampling for classification of imbalanced data is used in many applications. The feature-based similarity is used to generate synthetic instances among minority instances.

This method makes traditional classifiers to enhance the decision boundary close to minority instances.

3.1.1.2 ALGORITHM LEVEL APPROACH:

Traditional classifiers are modified to deal with imbalanced data.

3.1.1.3 RELATED WORK

There are Rapid technological inventions in the Data mining domain at an extraordinary pace. Its implementation of real-world problems in diverse areas - arises the problem of imbalanced nature of data. It has been considered as one of the Top 10 challenging problems in data mining [66]. Many researchers have accepted the challenge and proposed their solutions.
These solutions basically fall into two categories: Data level and Algorithmic level [67]. A book which is in a form of a paper collection edited by 'He and Ma' (2013) [68], It covers important issues such as sampling strategies, streaming data, and active learning. A book by García *et al*., [69] discussed data preprocessing steps such as preparing, cleaning and sampling imbalanced datasets. An in-depth insight into learning from skewed data and issues related to predictive modeling was discussed by Branco *et al.*, (2016) [70]. A more specialized discussion thorough a survey on ensemble learning is given by Galar et al., [71]. A global review on imbalanced data proposed by López *et al*., (2013) [72] and an in-depth discussion on new perspectives of evaluating classifiers on imbalanced datasets were presented [73]. A systematic review is done by Menardi and Torelli [74] on Class imbalance distribution. They have proposed a re-sampling method that leads to boosting and bagging to improve the accuracy of severe imbalanced data. Zhang and Li (2014) [75] performed experiments on three traditional classifiers, used mean and standard deviation to generate samples for the minority class. He stated that oversampling influences the performances of traditional classifiers.

3.1.1.4 EXPERIMENTAL INVESTIGATION

3.1.1.5 DATASETS

A study for binary class distribution on 12 data sets openly available with different degrees of imbalance nature is conducted. Table 1 contains the description of the data sets used for demonstrating the effectiveness of our proposed solution on various Parameters12 datasets that were used from the UCI or KEEL repository [76-77]. The number of instances, no. of attributes and degree of imbalanced distribution (imbalanced ratio) of the datasets are also given.

Table 3.1 Imbalanced Datasets with different degrees of Imbalanced distribution.

S. No.	Data Set Name	imbalanced Ratio	No. of Instances	No. of Attributes
1.	Abalone	129.44	4174	8
2.	Cleveland-0	12.62	177	13
3.	E. coli-3	8.6	336	7
4.	Glass-1	1.82	214	9
5.	Haberman	2.78	306	3
6.	New-Thyroid 1	5.14	215	5
7.	Page-Blocks 0	8.79	5472	10
8.	Pima	1.87	768	8
9.	Wine Quality White	58.28	1482	11
10.	Breast Cancer Wisconsin	1.86	683	9
11.	Yeast 1	2.46	1484	8
12.	Vowel	9.98	988	13

3.1.1.6 EXPERIMENT SETTING

The results are evaluated over 12 datasets with WEKA 3.6.9 and Orange 3.20 Data mining tools.

- **WEKA:** The WEKA workbench is a machine learning and data-preprocessing tool, under the GNU General Public License. WEKA, acronym is Waikato Environment for Knowledge Analysis, was developed at the University of Waikato in New Zealand. It is written in Java and can run on Linux, Windows, and Macintosh operating systems. The current stable version, 3.8.0, is compatible with Java 1.7 [78]. WEKA provides the support for the whole data mining process, viz., and preparation of the input data-by-data transformation and preprocessing, analyzing the data using learning schemes, and visualizing the data.

- **Orange:** Orange is a component-based data mining and machine learning software suite, it features visual programming for explorative data analysis that is existing in the front end and helps in visualization, libraries for scripting and Python bindings. Orange has widgets, supported on Mac OS, Windows and Linux platforms [79].

3.1.1.7 PROPOSED CLUSTER-BASED UNDER-SAMPLING

In Random Under-Sampling some instances are removed randomly [80]. So it is possible that the valuable instances may get thrown away which may contain potential information it results as inaccurate outcomes and predictions. The solution to this problem -is the integration of unsupervised learning with supervised learning. The main purpose of this method is to selectively discard majority instances from the datasets. Clustering algorithms group the similar characteristic instances in one cluster so it can have representation from the overall population. On the other hand Random over- Sampling instances are randomly replicated increasing the dataset size results in longer training time. It can be visualized using Fig. 4. Under-sampling [81] is a technique to reduce the number of samples in the majority class, where the size of the majority class sample is, reduced from the original datasets to balance the class distribution.

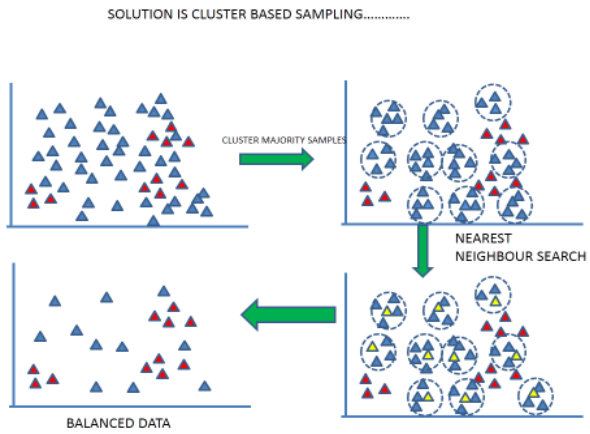


Figure 3.2 The framework of Cluster-Based sampling

3.1.1.8 METHODOLOGY

The overall process of transforming imbalanced data to balance can be divided into two phases: In the first phase, Between –Class imbalance and Degrees of imbalanced distribution of data were resolved. In the second phase, under-sampling using the clustering method will

be deployed. K- means algorithm to partition majority class instances into K groups will be used. The value of K will be optimized by the silhouette plot. To decide the imbalanced ratio for the resultant training set two ratios 0.50 and 1.0 will be used. then the number of instances to be selected from each cluster using equation-1 will be calculated. The required number of instances will be selected randomly from each cluster.

In order to evaluate the performance of the proposed method, the resultant training set will be applied to the SVM classifier and derive performance measuring parameters with IR = 0.50, IR = 1 and original dataset to identify the better performances.

Algorithm for balancing data using Clustering as an under-sampling tool:

Step 1: Segregate whole data set into MIN^{inst} and MAJ^{inst}

MIN^{size} –No. of instance belongs to Minority class

MAJ^{size} –No. of instances belongs to Majority instances.

In imbalanced datasets$MAJ^{size} > MIN^{size}$

Step 2: Removing Outliers

MIN_i^{inst} where i=1, 2, 3, 4…….MIN^{size}

MAJ_j^{inst} where j=1, 2, 3, 4……..MAJ^{size}

If Distance (MIN_i^{inst} ,MAJ_j^{inst}) =0

Remove MAJ_j^{inst} from MAJ^{size} and update MAJ^{size}

Step 3: Decide the imbalanced ratio for each cluster by setting the ratio parameter from IR= {0.50, 1}

Step 4: Build clusters from majority instances using K mean algorithm. Draw a silhouette plot to find the most appropriate value of K.

$$MAJ^{size} = \sum_{i=1}^{k} C_i^{inst}$$

Step 5: NC^i no. of instances in Ith cluster

SC_i^{inst} No. of instances to be sampled from each cluster is defined as:

$$SC_i^{inst} = IR \times \frac{MIN^{size}}{MAJ^{size}} \times NC^i$$

Step 6: Repeat the process to find no. of instances to be selected from each cluster no for i=1, 2, 3, 4……..….k

Step 7: Randomly select any instance from a given cluster

If Distance (C_1^i,C_{i+1}^i) =0

Add C_1^i in the sampled training set and

Remove C_1^i and duplicated C_{i+1}^i from cluster C_i^{inst}

Repeat the process until we get the required instances from the cluster or instances of the cluster to get exhausted.

Step 8: To get a balanced cluster, all output instances from each cluster with Minority instances will be merged to get a final Balanced training set.

Fig. 3.5 represents the flow control of the proposed model. The classifier takes minority class instances as noise and does not consider them in the building model so the classifier gets biased towards the majority class. The Class imbalance, class overlap added with high dimension data makes classifying tasks complicated and challenging.

As stated above, this model is capable of giving 4 fold solutions. The first solution gives the capability of handling the different degrees of imbalanced nature of data. In this approach, the majority and minority instances are separated into two datasets and majority class instances are handled separately so in the first phase only we can handle the diverse degree of imbalanced distribution. It balances imbalanced data using an under-sampling method. Here in this approach, a cluster-based method is deployed to reduce the size of the majority class to make training set balanced. It is capable to handle between class imbalanced and within-class imbalanced nature of data. Between classes, imbalance distributions are solved in the first phase when majority and minority instances are classified into two sets and within-class imbalanced problems can be solved by making clusters from majority class and selecting uniform cluster representatives from each cluster. The outcomes of the experiments conducted on 12 datasets proved how the proposed algorithm reduces Type-1 (False positive) and Type-2 (False negative) errors which are a very serious concern while working on medical sophisticated datasets.

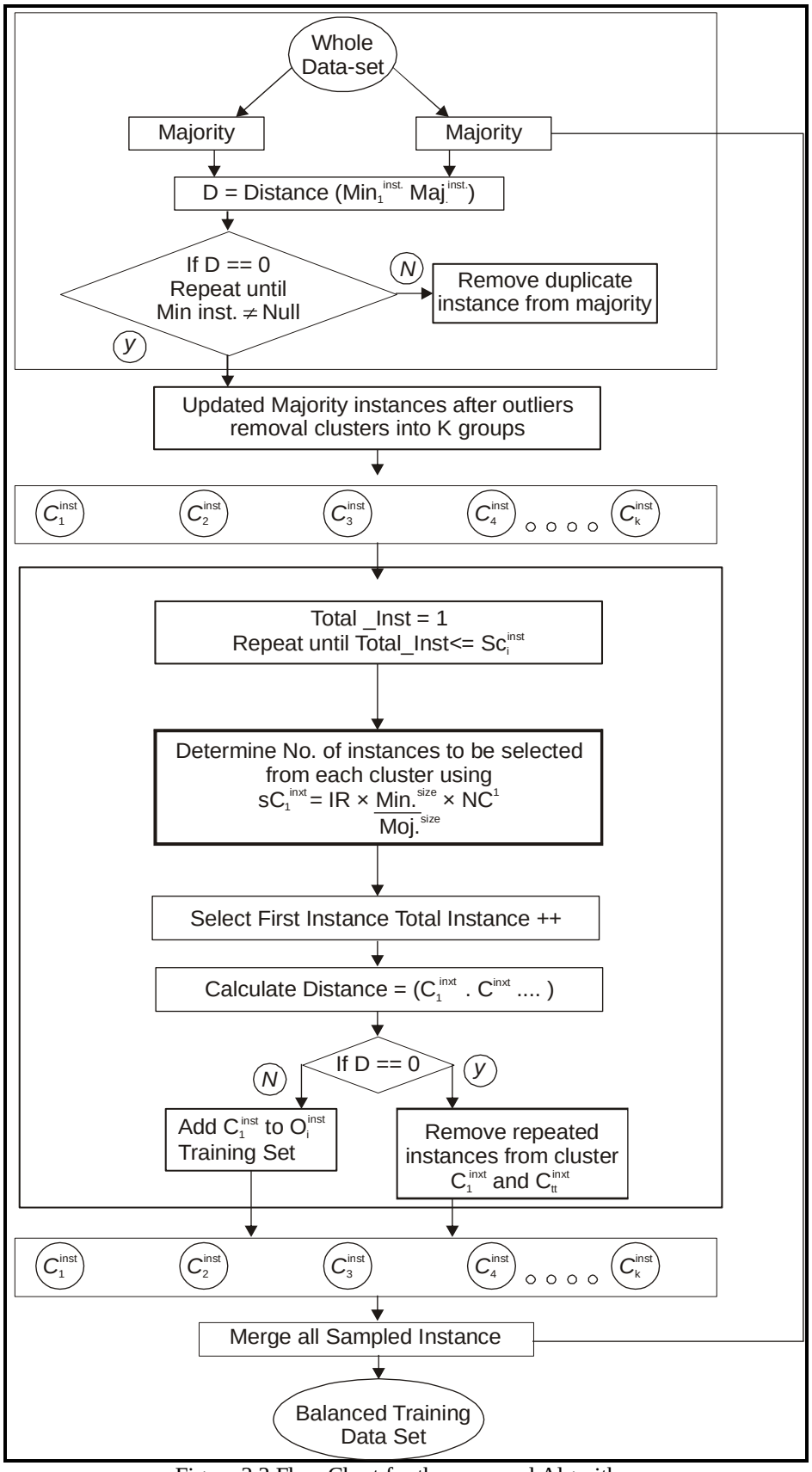


Figure 3.3 Flow Chart for the proposed Algorithm

Table 3.2 Overall performance of the model on various parameters

S. No	Data set name	Original data set					Balanced with IR = 0.50					Balanced with IR =1				
		Acc	F-mea	FP	Pre	Roc	Acc	F-mea	FP	Pre	Roc	Acc	F-mea	FP	Pre.	Roc
1.	Abalone	98	0.89	0.16	0.90	0.90	97	0.96	0.07	0.96	0.98	97	0.93	0.58	0.94	0.97
2.	Cleveland-0	95.9	0.97	0.38	0.9	0.80	80	0.84	0.23	0.84	0.78	96.1	0.98	0.00	1.0	0.96
3.	E coli-3	89.5	0.94	1.0	0.89	0.50	92.3	0.94	0.17	0.91	0.89	81.1	0.83	0.32	0.75	0.81
4.	Glass-1	63.5	0.77	1.0	0.64	0.49	78	0.86	0.59	0.77	0.74	70	0.78	0.54	0.64	0.70
5.	Haberman	73	0.84	1.0	0.73	0.50	66.9	0.80	1.0	0.66	0.50	57	0.46	0.2	0.64	0.57
6.	New-thyroid 1	92	0.95	0.45	0.91	0.77	90.1	0.93	0.31	0.87	0.84	98	0.98	0.03	0.97	0.98
7.	Page-blocks 0	93	0.93	0.59	0.93	0.70	84	0.88	0.18	0.90	0.83	85	0.95	0.08	0.90	0.85
8.	Pima	77.3	0.62	0.10	0.74	0.72	77.3	0.84	0.49	0.78	0.70	71.7	0.70	0.04	0.73	0.75
9.	Wine quality white	98.3	0.99	1.0	0.98	0.50	62	0.76	1.0	0.65	0.46	69.3	0.70	0.33	0.69	0.69
10.	Breast cancer Wisconsin	96.9	0.97	0.03	0.98	0.96	98.3	0.99	0.02	0.99	0.78	97	0.97	0.03	0.97	0.77
11.	Yeast 1	74.3	0.84	0.81	0.74	0.57	80	0.89	0.23	0.83	0.80	76	0.85	0.31	0.75	0.68
12.	Vowel	95.9	0.73	0.50	0.91	0.80	90	0.84	0.11	0.92	0.93	96	0.70	0.23	0.89	0.74

Table 3.2 presents a comparative analysis of the performance of the SVM classifier on original imbalanced data of different degrees, Data with imbalanced ration = 0.50 and data with imbalanced ratio = 1. It also presents collective information in order to identify which method's performance is better over others. Fig. 3.4 is a pictorial representation of the accuracy of the SVM classifier for an original dataset, with an imbalanced Ration of 0.50 and an imbalanced ration of 1 on 12 datasets. As we have discussed accuracy is not the perfect measure for imbalanced data.

In the graph, high accuracy with few original imbalanced data sets is achieved but it does not indicate a better overall performance as accuracy is not the only measure to check the performance of the classifier. Fig. 3.5 is a pictorial representation of the F-measure of SVM classifier for the original dataset, with an imbalanced Ration of 0.50 and an imbalanced ration of 1 on 12 datasets. It can be easily identified that we are getting improved values for the FP rate with the proposed method.

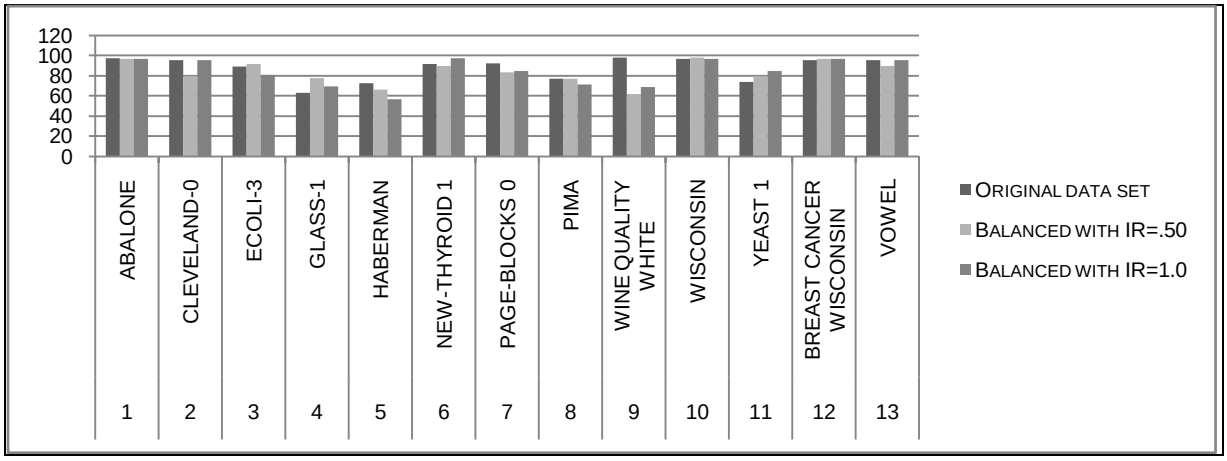


Fig. 3.4 Chart for accuracy.

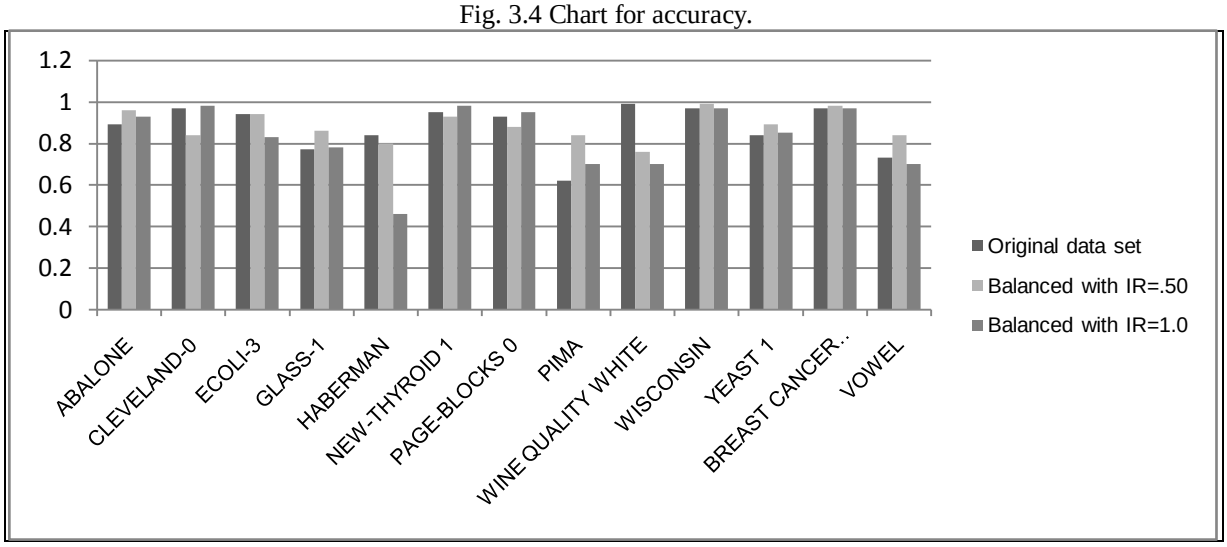


Fig. 3.5 Chart for F-Measure

Fig. 3.6 is a pictorial representation of the F-P rate of SVM classifier for an original dataset, with an imbalanced Ration of 0.50 and an imbalanced ration of 1 on 12 datasets.

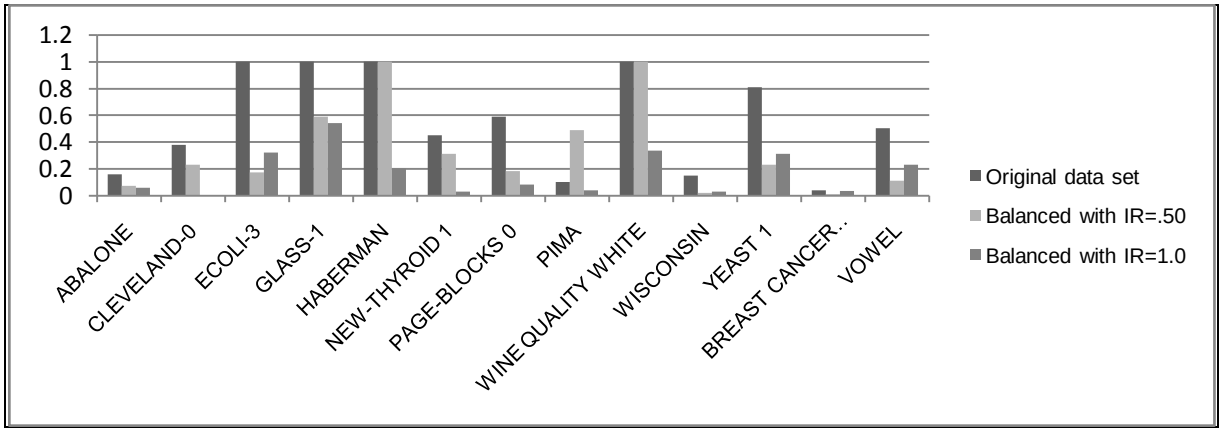


Fig. 3.6 Chart for F-P rate

3.7 is a pictorial representation of Precision of SVM classifier for an original dataset, with imbalanced Ration of 0.50 and imbalanced ration of 1 on 12 datasets

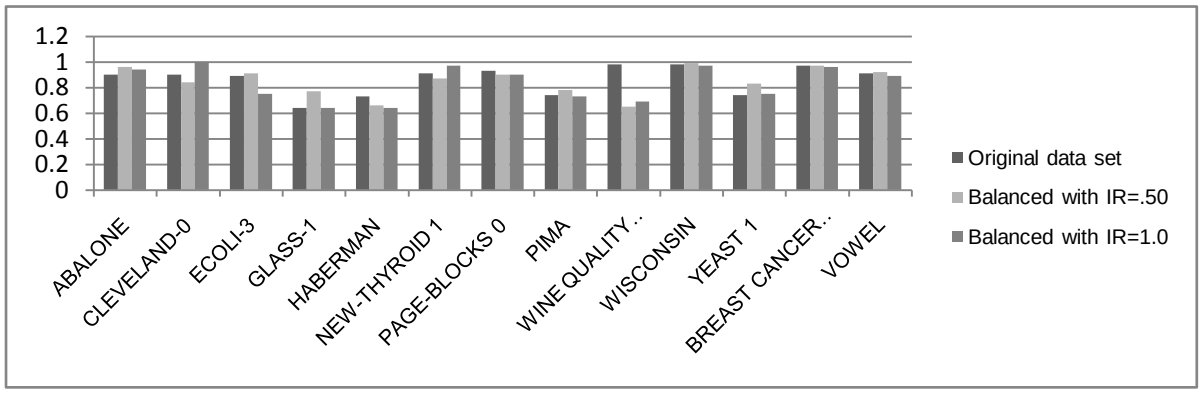


Fig. 3.7. Chart for precision

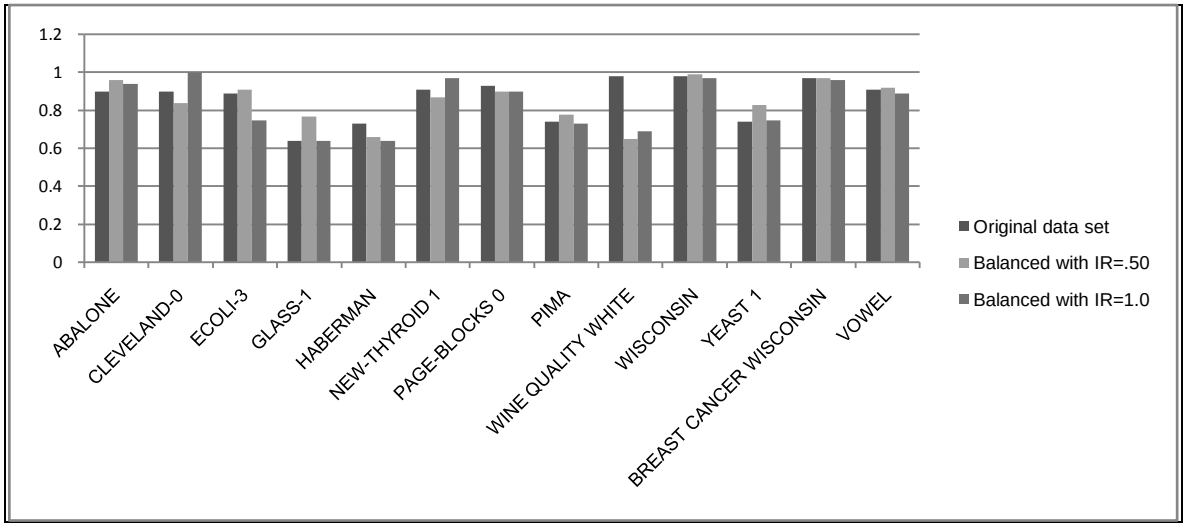


Fig. 3.8. Chart for ROC.

Fig. 3.8 is a pictorial representation of ROC of SVM classifier for an original dataset, with an imbalanced Ration of 0.50 and an imbalanced ration of 1 on 12 datasets.

3.2 CONCLUSION

In this research work, the cluster-based under-sampling method is applied on a different degree of imbalanced ratio over 12 data sets from UCI and KEEL repositories. Experimental results show the better performance of the proposed algorithm on these data sets. It also deals with the two significant problems while working on imbalanced data they are between- class imbalance distribution of data and within-class imbalance distributions of data.

CHAPTER 4

4 A HYBRID MODEL TO ENHANCE THE PERFORMANCE OF A CLASSIFIER

4.1 INTRODUCTION

The complexity and volume of data are increasing day by day; the existing data mining techniques are facing many challenges particularly for classifying large-scale multi-class data. Therefore, the combination of clustering and classification becomes an active research area to deal with data in large volumes with high dimensions.

Data mining problems are being solved by Single model method (supervised or unsupervised techniques), Ensemble methods ((combine homogeneous machine Learning approaches) or Hybrid methods (combine heterogeneous machine Learning approaches).

If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model. The objective is to utilize the strength of one method to complement the weaknesses of another.

4.2 DATASETS DESCRIPTION

The experiment conducted on 13-benchmark dataset taken from the UCI machine-learning repository.

Table 4.1 Data set description

S. No	Dataset Name	#Features	#Instances	Class
1	Heart Disease	14	303	2{Yes, No}
2	Wine quality white	12	1482	2{Negative, Positive}
3	Parkinson dataset	756	757	2{Yes, No}
4	Colon Cancer	2002	62	2{Normal, Abnormal}
5	Breast Cancer	32	569	2{Malignant, Benign}
6	Image Segmentation	19	210	7{Brickface, Sky, Foliage, Cement, Window, Path, Grass}
7	Cleveland-0 Dataset	13	173	2{Negative, Positive}
8	Ionosphere Dataset	35	351	2{Good, Bad}
9	Squash Harvest Stored	25	53	3{Excellent, Ok, Not acceptable}
10	Bank Dataset	17	4522	2{Yes, No}
11	Glass Dataset	10	215	2{Positive, Negative}
12	Pima	9	769	2{Positive, Negative}
13	Haberman Dataset	4	307	2{Positive, Negative}

4.3 METHODOLOGY

4.3.1 DATA PREPARATION

During the initial phase separate the datasets into a training set and testing sets and remove output classes from both training and testing datasets. Removing output class labels is important to make the clusters unbiased of class attributes.

4.3.2 CLUSTERS BUILDING

Prior to the classification, the dataset is grouped into the different number of clusters using the K- means algorithm depending upon the value of K (K=2, 3, 4 and 5). When k= 2, two clusters are formed from the training set and two clusters from a testing set. When k=3, three clusters are formed from training and three from testing and when k=4, four clusters for training and four from testing and same for k=5 after this whole process, fourteen clusters $\{ I_1^{k^2}, I_2^{k^2}, I_1^{k^3}, I_2^{k^3}, I_3^{k^3}, I_1^{k^4}, I_2^{k^4}, I_3^{k^4}, I_4^{k^4}, I_1^{k^5}, I_2^{k^5}, I_3^{k^5}, I_{4,}^{k^5} I_5^{k^5} \}$ for testing and fourteen $\{I_1^{k^2}, I_2^{k^2}, I_1^{k^3}, I_2^{k^3}, I_3^{k^3}, I_1^{k^4}, I_2^{k^4}, I_3^{k^4}, I_4^{k^4}, I_1^{k^5}, I_2^{k^5}, I_3^{k^5}, I_{4,}^{k^5} I_5^{k^5}\}$ for training without output classes are formed. Corresponding output classes are added to each cluster. After this process, the Classification task can be applied. This process is being performed to identify which value of K is most suitable for which classification algorithm given dataset.

4.3.3 BUILDING THE CLASSIFICATION MODELS*:*

The resultant dataset augmented with one more feature originated from the clustering step named cluster_no as an input to the classifier. The expanded data is applied to SVM (Support Vector Machine) classifier. This model will be evaluated on several parameters. The results evidently revealed that this integration process generates a more precise and accurate model. All classifier's accuracy gets affected positively when supervised and unsupervised learning methods are combined.

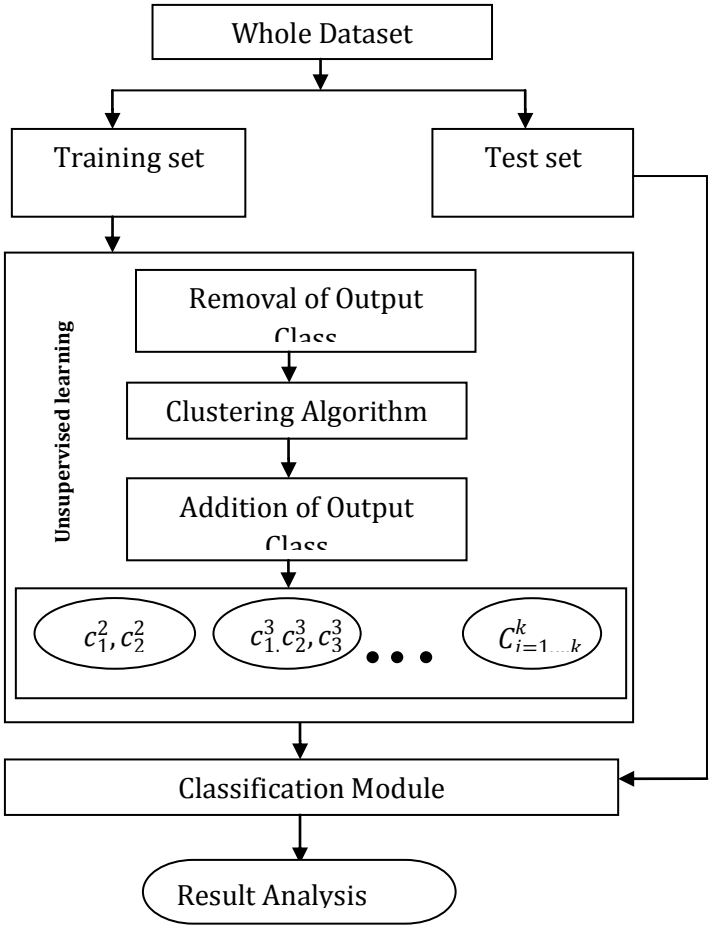


Figure 4.1 Framework for the hybrid model

Table 4.2 Proposed model performance on different cluster number

s. no	Dataset Name	Whole Dataset without Clustering			Clustered Data set k=2			Clustered Dataset k=3			Clustered Dataset k=4			Clustered Dataset k=5		
		F-M	ROC	Acc.	F-M	ROC	Acc	F-M	ROC	Acc.	F-M	ROC	Acc.	F-M	ROC	Acc.
1	Heart Disease	.85	.82	83.4	.99	.99	99.6	.97	.98	98.6	.97	.99	98.0	.98	.98	98.3
2	Wine-quality	.99	.98	98.3	.99	.50	98.3	.99	.50	98.3	.95	.98	96.6	.99	.50	98.3
3	Parkinson	.90	.78	85.7	.90	.78	85.4	.91	.80	87.5	.91	.80	86.7	.91	.80	87.0
4	Colon	.88	.83	85.4	.88	.83	85.4	.88	.83	85.4	.88	.83	85.4	.85	.78	80.6
5	Breast cancer	.97	.97	97.8	.99	.99	99.8	.99	.99	99.8	.99	.99	99.8	.99	.99	99.8
6	Image	.80	.96	88.5	.96	1.0	95.2	.96	.99	94.7	.91	.97	94.2	.95	.99	96.6
7	Cleland	.97	.80	95.9	.98	.84	97.1	.98	.84	97.1	.98	.84	97.1	.98	.84	97.1
8	Ionosphere	.82	.85	88.6	1	1	100	1	1	100	1	1	100	1	1	100
9	Squash	.76	.83	71.1	.88	.88	80.7	.95	.95	90.3	.92	.93	86.5	.90	.92	88.4
10	Bank data	.94	.57	89.2	.94	.57	89.2	.94	.57	89.2	.94	.57	89.2	.95	.66	90.9
11	Glass	.77	.49	63.5	.76	.48	62.1	1	1	100	1	1	100	.99	.99	99.5
12	Pima	.62	.72	77.3	1	1	100	1	1	100	1	1	100	1	1	100
13	Haberman	.84	.50	73.5	.84	.49	73.2	.84	.49	73.2	.84	.50	73.5	1	1	100

4.4 ALGORITHM

Step 1: Preprocessing the dataset by deleting all missing values instances

D= $\{I^{\text{instance}}, O^{\text{instance}}\}$

Here I stand for Input Instance and O stands for output instances

Step 2: Divide 2/3 of the dataset as training and 1/3 as testing

TrainingSubset =2/3$\{ I^{\text{instance}}, O^{\text{instance}} \}^{\text{training}}$

TestingSubset = 1/3 $\{ I^{\text{instance}}, O^{\text{instance}} \}^{t\text{esting}}$

Step 3: Remove the Corresponding output class from the dataset

$I_1^{k^2}$TrainingSubset = $\{ I^{\text{instance}} \}^{\text{training}}$

Step 4: Apply K- means clustering algorithm for k=2, 3,4 and 5 cluster

TrainingSubsets = $\{I_1^{k^2}, I_2^{k^2}, I_1^{k^3}, I_2^{k^3}, I_3^{k^3}, I_1^{k^4}, I_2^{k^4}, I_3^{k^4}, I_4^{k^4}, I_1^{k^5}, I_2^{k^5}, I_3^{k^5}, I_{4,}^{k^5} I_5^{k^5}\}$

TestingSubsets = $\{I_1^{k^2}, I_2^{k^2}, I_1^{k^3}, I_2^{k^3}, I_3^{k^3}, I_1^{k^4}, I_2^{k^4}, I_3^{k^4}, I_4^{k^4}, I_1^{k^5}, I_2^{k^5}, I_3^{k^5}, I_{4,}^{k^5} I_5^{k^5}\}$

Step 5: Add target classes for each cluster

TrainingSet 1 = $\{ \{C_1^{k^2} + O_{\text{Training}}^{\text{instance}}\}$

TrainingSet 2 = $\{ \{C_2^{k^2} + O_{\text{Training}}^{\text{instance}}\}$

Step 6: Train classifier by the clustered datasets for each combination of K.

Step 7: for each value of K get the corresponding constructed model name $M_n^{k^i}$ where n defines model number and k^i defines cluster number

Step 8: Integrate all cluster values classified to produce the outcome for the whole dataset.

Step 9: Create a confusion matrix for furthest analysis

4.5 RESULTS & CONCLUSION

Figure-4.2 gives a comparative sight of unprocessed data with K= 2, 3, 4 and 5. The plot displays the F measure, Accuracy and ROC area. It can be identified by seeing the figure that we get an elevated performance with a clustered dataset on various performance measuring parameters.

All classifier's accuracy gets affected positively when supervised and unsupervised learning methods are combined.

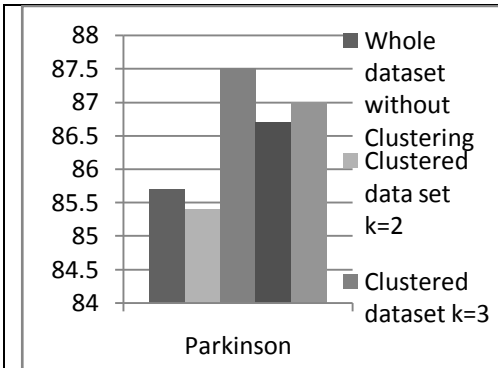


Figure 2 Chart for Accuracy on different cluster no for Parkinson Dataset

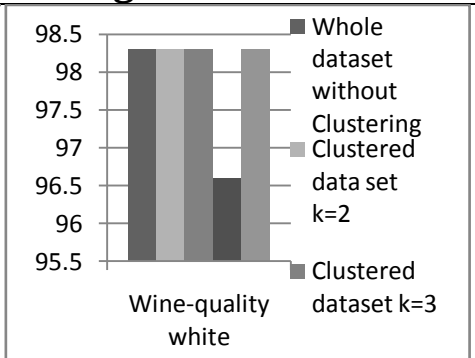


Figure 3 Chart for Accuracy on different cluster no for Wine Quality dataset

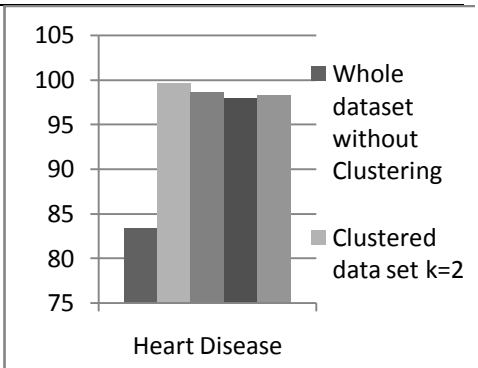


Figure 4 Chart for Accuracy on different cluster no for Heart disease Dataset

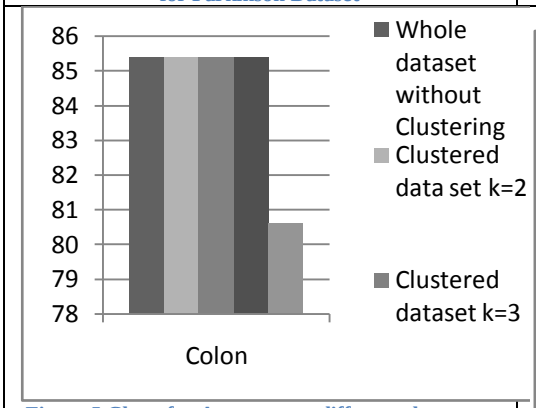


Figure 5 Chart for Accuracy on different cluster no for Colon Dataset

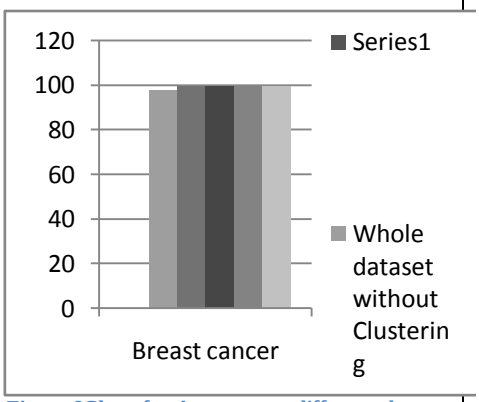


Figure 6Chart for Accuracy on different cluster no for Breast Cancer dataset

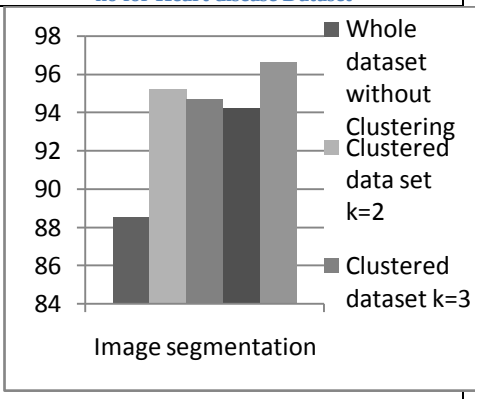


Figure 7 Chart for Accuracy on different cluster no for Image Segmentation dataset

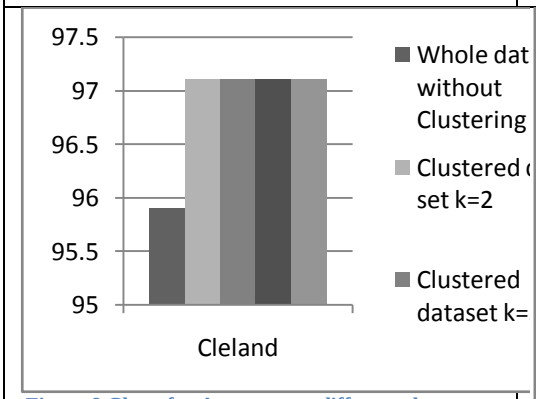


Figure 8 Chart for Accuracy on different cluster no for Cleland dataset

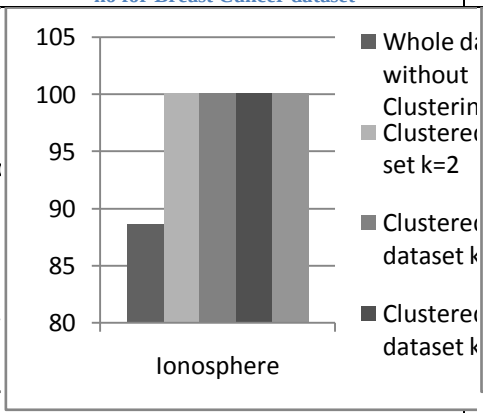


Figure 9 Chart for Accuracy on different cluster no for Ionosphere dataset

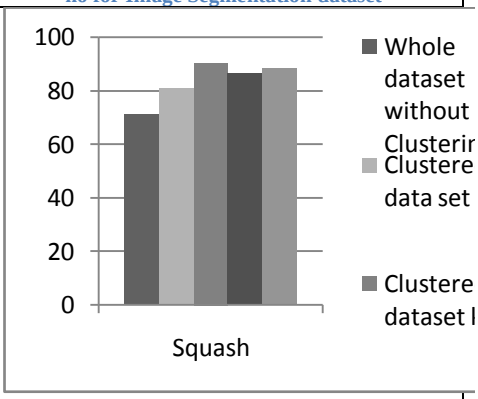


Figure 10Chart for Accuracy on different cluster no for Squash dataset

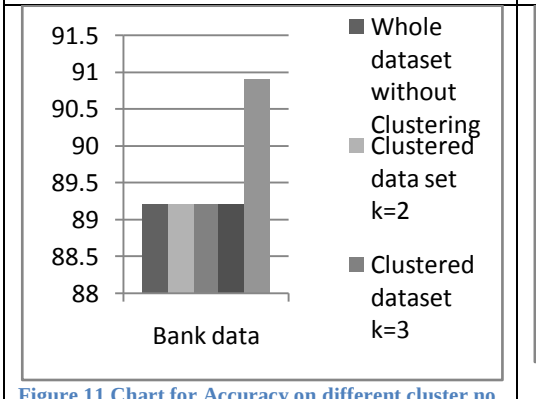


Figure 11 Chart for Accuracy on different cluster no for Bank Dataset

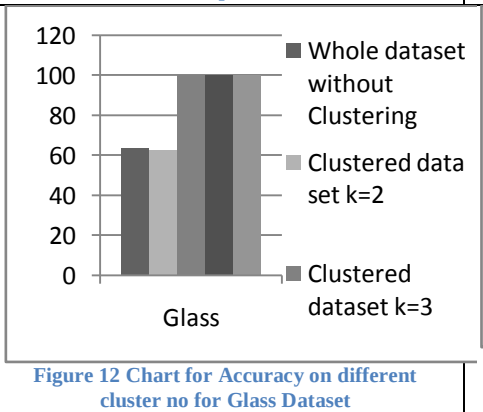


Figure 12 Chart for Accuracy on different cluster no for Glass Dataset

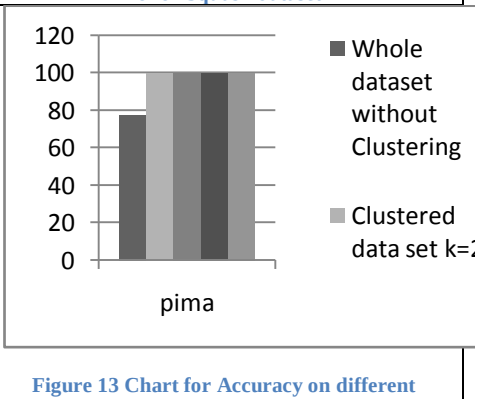


Figure 13 Chart for Accuracy on different cluster no for Pima Dataset

Figure 4.2: Charts for the Accuracy of classifiers on five values of K for different datasets

Chapter 5

5 CLUSTERING-BASED FEATURE SELECTION METHOD

5.1 INTRODUCTION

Innovative computer technologies are omnipresent. Data is accumulating with an unmatchable speed of the human's capacity to process on it. High dimensional datasets suffer from problems such as bogus correlations, and heavy computation cost. Multi-dimensional data diminish the performance of Data Mining algorithms and increases computation cost involved in processing. Irrelevant features sometimes exert an undue load on classifiers and increase the time and resources needed to build the model. Moreover, high dimensional data sets may contain groups of correlated attributes they measure the same underlying meaning. The irrelevant dataset can also mislead the logic of the algorithm that affects the performance of the model. The proposed model identifies the relevant features that may lead to accurate results.

A better feature set can solve numerous machine-learning problems. There are two commonly used approaches to reduce the dimensionality of the feature set. The first one is Feature extraction where we transform the existing features into a less dimensional space and the second one is the Feature selection where we select a subset from the original features space without transforming them. Feature selection /extraction is a proven pre-processing step for reducing dimensionality, increasing comprehensibility, improving the overall performance and to reduce the complexity and computation time of any classification algorithms. The main purpose of the proposed method is to select the most influencing relevant feature space by discarding irrelevant features. The purpose of feature subset selection is to select and remove as many redundant and irrelevant features as possible.
There are a few significant works done for feature selection and Extraction. Little et al(2008) [84] proposed a new measure named Dysphonia , PPE(pitch period entropy) and used Support Vector Machine(Gaussian radial basis kernel functions) to classify PD dataset and reported 91.4% of accuracies and Das discussed a Neural Network classification scheme and reported 92.9% as percentage accuracies. Luukka (2011)[85] introduced a fuzzy entropy-based feature selection method to predict PD. The reported classification accuracy was 85.03% with only two features from the original set of features. Li, Liu, and Hu (2011) [86] proposed a combination of a fuzzy-based non-linear transformation method with the principal component analysis (PCA) in order to extract the optimal set of features for SVM classifier. The best-reported classification accuracy of 93.47% was achieved.

INTRODUCTION ABOUT DATA SET
Data set from the UC Irvine Machine Learning Repository [82] viz. Heart Disease, Wine quality white, Parkinson dataset, Colon Cancer, Breast Cancer, Image Segmentation, Cleveland-0, Ionosphere, and Squash Harvest Stored are used for the study.

Table 5.1 Data set description

S. No	Dataset Name	#Features	#Instances	Class
1	Heart Disease	14	303	2{Yes, No}
2	Wine quality white	12	1482	2{Negative, Positive}
3	Parkinson dataset	756	757	2{Yes, No}
4	Colon Cancer	2002	62	2{Normal, Abnormal}
5	Breast Cancer	32	569	2{Malignant, Benign}
6	Image Segmentation	19	210	7{ Brickface, Sky, Foliage, Cement, Window, Path, Grass}
7	Cleveland-0	13	173	2{Negative, Positive}
8	Ionosphere	35	351	2{Good, Bad}
9	Squash Harvest Stored	25	53	3{Excellent, Ok, Not acceptable}

5.2 Methodology

Co-relation based clustering algorithms are not only applicable to cluster data based on their similarity but also useful for feature compression, and reduction technique. The correlation-based feature selection method(K- means) has been proposed and applied to classify the diversified dataset.

The diagnosis process involves four steps: data pre-processing, feature extraction /selection to identify and remove irrelevant, redundant, or noisy features from the provided dataset, data classification and performance evaluation. The reduced dimensional feature set is used as input to each of the classifiers.

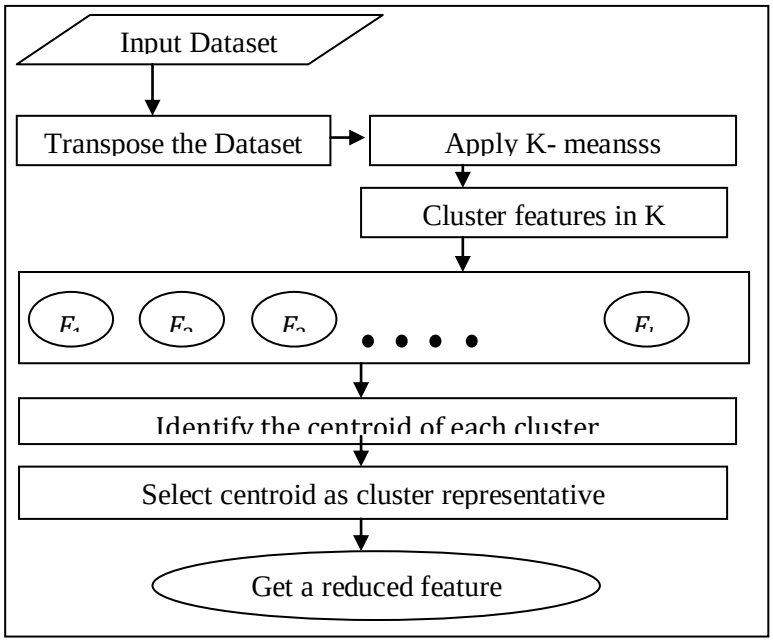


Figure 5.1 Flow chart for the proposed model

5.3 SVM Classifier:

The support vector machine(SVM) is the most appropriate, effective and popular non-linear statistical learning method among all classification algorithms. It is especially applicable for diagnosis and prognosis classification problems because of its high generalization capacity. [89]

Table 5.1 Comparative analysis of the Proposed Approach with Relief & Info-Gain Feature Selection approach

Dataset	% of feature selected	No of features selected	Proposed Approach: (Clustering-based Feature Selection Model)			RELIEF: Feature Selection Approach			Info-Gain: Feature Selection Approach		
			F-Measure	ROC Area	Accuracy	F-Measure	ROC Area	Accuracy	F-Measure	ROC Area	Accuracy
Heart Disease Dataset	20 %	3	.73	.69	70.0	.79	.77	77.5	.83	.80	81.1
	40 %	6	.79	.76	76.8	.81	.78	78.8	.85	.82	83.1
	60 %	8	0.96	0.96	96.3	.84	.81	82.1	.85	.82	83.1
	80 %	11	.81	.78	79.2	.85	.82	82.8	.85	.82	83.4
	Whole Data set	14	.85	.82	83.4	.85	.82	83.4	.85	.82	83.4
Wine-quality white Dataset	20 %	3	.99	.98	98.3	.97	.97	97.8	.99	.50	98.3
	40 %	5	.99	.98	98.3	.99	.98	98.3	.99	.50	98.3
	60 %	8	.99	.98	98.3	.99	.98	98.3	.99	.50	98.3
	80 %	10	.99	.98	98.3	.99	.98	98.3	.99	.50	98.3
	Whole Data set	12	.99	.98	98.3	.99	.98	98.3	.99	.98	98.3
Parkinson Dataset	20 %	151	.88	.74	85.0	.92	.79	85.5	.91	.76	86.3
	40 %	302	.88	.74	86.2	.91	.79	86.1	.91	.78	86.1
	60 %	454	.89	.78	86.4	.91	.78	86.1	.91	.78	86.0
	80 %	605	.90	.78	85.9	.90	.77	85.4	.90	.78	85.4
	Whole Data set	756	.90	.78	85.7	.90	.78	85.7	.90	.78	85.7
Colon Dataset	20 %	400	.88	.83	85.4	.84	.75	79.0	.84	.75	79.0
	40 %	800	.88	.83	85.4	.85	.76	80.6	.85	.76	80.6
	60 %	1201	.88	.83	85.4	.86	.79	82.2	.85	.76	80.6
	80 %	1602	.88	.83	85.4	.86	.79	82.2	.86	.79	82.2
	Whole Data set	2002	.88	.83	85.4	.88	.83	85.4	.88	.83	85.4
Breast cancer Dataset	20 %	7	.90	.92	93.1	.94	.95	96.3	.91	.93	94.0
	40 %	13	.92	.93	94.5	.95	.95	96.8	.91	.92	93.6
	60 %	19	.95	.95	97.6	.96	.96	96.5	.95	.95	96.6
	80 %	26	.96	.96	97.3	.97	.97	97.8	.97	.97	97.8
	Whole Data set	32	.97	.97	97.8	.97	.97	97.8	.97	.97	97.8
Image segmentation Dataset	20 %	4	.47	.84	57.6	.28	.74	49.0	.37	.76	50.0
	40 %	8	.65	.91	75.2	.48	.84	70.4	.82	.99	72.3
	60 %	11	.73	.95	95.6	.79	.96	87.6	.91	.99	90.0
	80 %	15	.69	.92	85.7	.78	.95	89.0	.87	.99	87.6
	Whole Data set	19	.80	.96	88.5	.80	.96	88.5	.80	.96	88.5
Cleland Dataset	20 %	3	.96	.50	92.4	.97	.76	94.7	.96	.57	93.6
	40 %	6	.96	.68	93.6	.97	.68	94.7	.97	.76	95.9
	60 %	9	.98	.80	96.5	.97	.76	95.3	.97	.76	95.3
	80 %	11	.98	.80	96.5	.97	.80	95.9	.97	.76	95.3
	Whole Data set	14	.97	.80	95.9	.97	.80	95.9	.97	.80	95.9
Ionosphere Dataset	20 %	7	.89	.82	86.0	.87	.77	82.3	.87	.77	82.3
	40 %	14	.90	.83	86.6	.89	.82	86.0	.89	.82	86.0
	60 %	21	.89	.82	85.7	.88	.78	83.4	.88	.78	83.4
	80 %	28	.90	.82	86.3	.90	.82	86.3	.88	.78	83.4
	Whole Data set	35	.82	.85	88.6	.82	.85	88.6	.82	.85	88.6
Squash Dataset	20 %	5	.57	.61	50	.73	.74	59.6	.73	.77	59.6
	40 %	10	.65	.72	61.5	.69	.74	57.6	.69	.71	57.6
	60 %	15	.73	.78	63.4	.64	.71	53.8	.77	.81	67.3
	80 %	20	.73	.78	63.4	.76	.84	73.0	.73	.79	67.3
	Whole Data set	25	.76	.83	71.1	.76	.83	71.1	.76	.83	71.1

5.4 PERFORMANCE ANALYSIS OF THE GENERATED MODEL

The objective of this experiment is to minimize the feature set & to maximize the accuracy of the classifier on those reduced feature set. Thus in this experiment, a different percentage of features are selected and tested on some standard performance measuring parameters such as F-Measure, ROC and Accuracy. There are three methods discussed in the literature survey to select cluster representative they are, Random feature selection, the top-ranked feature from each cluster or cluster centroid from each cluster. In this study, cluster centroid is used to select a cluster representative.

Chart for Heart disease dataset

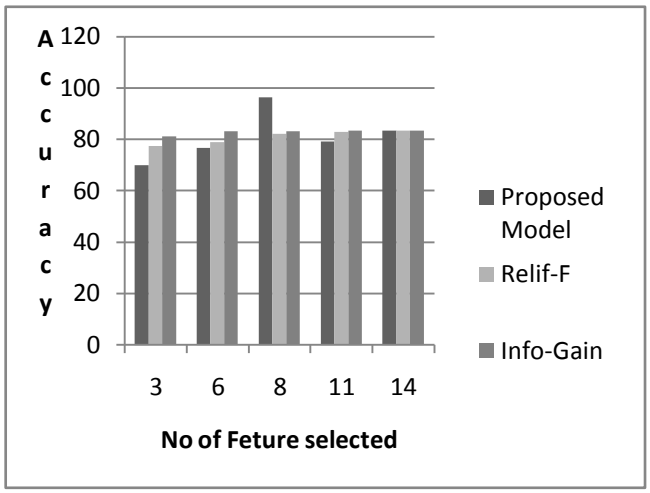


Chart for colon cancer dataset

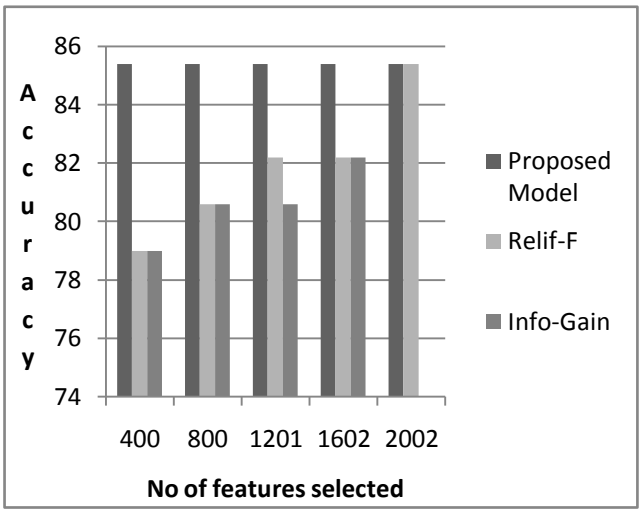


Chart for Image segmentation

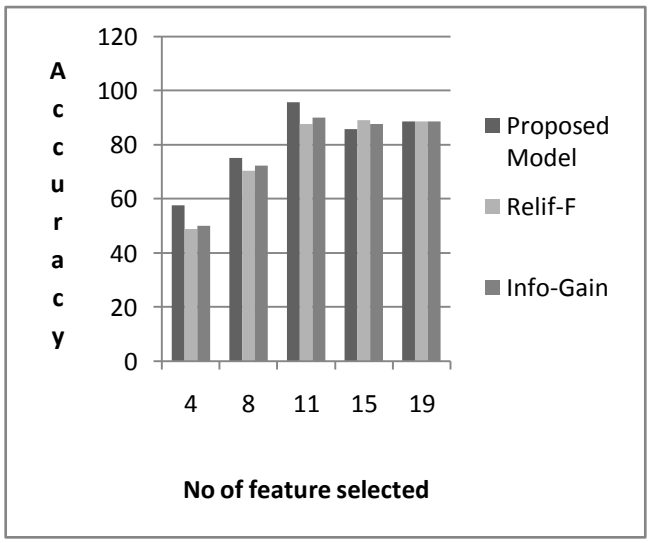
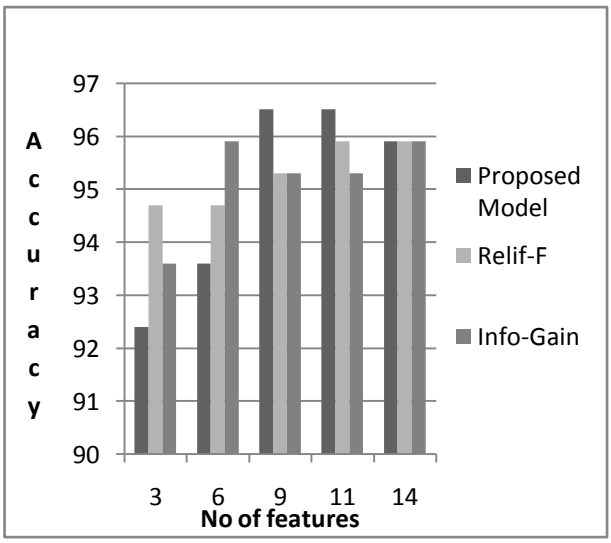


Chart for Cleland dataset



5.5 Results & Conclusion

Table 5.1 gives a comparative analysis of the Proposed Approach with a standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) on the aforementioned nine datasets. The analysis is presented on 20, 40, 60, 80 and 100 % of the features on 3 performance-measuring parameters Accuracy, F-Measure and ROC. Figure 5.2 to 5.5 presents a plot on the accuracy of SVM on a different set of features for 5 datasets. The results revealed that this proposed method is simple yet effective to improve the performance of the classifier on a reduced feature set.

5.6 Further Enhancements

The proposed framework is capable to give an efficient model to deal with four machine learning problems viz. a feature compression & extraction technique, a fully labeled training set from the unlabeled set, to improve the performance of the classifiers, and to balance data from imbalanced data. only three models are proposed in this present research work. The work can be extended for achieving a fully labeled data set that can be achieved using a proposed integrated framework.

In chapter three experiments are carried out for binary classification, which could be extended for multiclass classification. In the fourth chapter, the models are tested on the SVM classifier, can be also be examined on other classifiers. In the fifth chapter high dimensional datasets can also be considered.

6 REFERENCES

[1] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus Usama Fayyad, Gregory Piatetsky-Shapiro ,Padhraic Smyth," Knowledge Discovery and Data Mining: Towards a Unifying Framework", in KDD-96 Proceedings. 1996.
[2] J.Han, J.Pei,and M. Kamber, "Data Mining: Concepts and techniques" in Elsevier,2011.
[3] M. H. Dunham, Data Mining, "Introductory and advanced topics" Pearson Education India, 2006.
[4] Y. Fu," Data Mining," IEEE potentials,vol.16, pp.18-20,1997.
[5] M. Durairaj, V. Ranjani ,"Data Mining Applications, In Healthcare Sector "International Journal of Scientific & amp; Technology Research, October 2013 ISSN 2277-8616 29.
[6] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth ,"The KDD process for extracting useful knowledge from volumes of data & quot;, Communications of the ACM 39.11 (1996): 27-34.
[7] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview & quot;, Advances in Knowledge Discovery and Data Mining,The MIT Press, Menlo Park, CA, 1996, pp.1-34.
[8] Masand, B. and G. P. Shapiro" A comparison of approaches for maximizing business payoff of prediction models",In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon, USA, pp. 195–201.
[9] Pang-Ningtan, vipinkumar, Michael Steinbach, " introduction to Data-mining" pearson 2008.
[10] Duda, R. O., Hart, P.E., and Stork, D.G.," Pattern Classification", John Wiley & Sons Inc., USA,2001
[11] G K, G. Kesavaraj & Sukumaran, Surya, "A study on classification techniques in data mining", 4th International Conference on Computing, Communications and Networking Technologies, 2013.p.p 1-7.
[12] Xinyang Deng, Qi Liu, Yong Deng, Sankaran Mahadevan , "An improved method to construct basic probability assignment based on the confusion matrix for classification problem", Information Sciences, Volumes 340–341, 2016, Pages 250-261, ISSN 0020-0255.
[13] Höppner, F., Klawonn, F., Kruse, R., and Runkler, T.," Fuzzy Cluster Analysis", Chichester John Wiley &Sons, 2000
[14] A. K. Jain and R. C. Dubes," Algorithms for Clustering Data",Prentice Hall, New Jersey, 1988.
[15] Jain AK, Murty MN, Flynn PJ," Data Clustering: A Review", ACM Computing Surveys, 31:264-323, 1999.
[16] Mduma, N., Kalegele, K. and Machuve, D., 2019. A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. Data Science Journal, 18(1), p.14.
[17] Feng Chen,Pan Deng, Jiafu Wan, Daqiang Zhang, AthanasiosV. Vasilakos, and Xiaohui Ron, Review Article Data Mining for the Internet of Things: Literature Review and Challenges Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2015, Article ID 431047, 14 pages
[18] DongkuanXu ,YingjieTian,"A Comprehensive Survey of Clustering Algorithms", Ann. Data. Sci. (2015),Springer-Verlag Berlin Heidelberg 2015, 165–193.
[19] Amatul Zehra, Tuty Asmawaty, M.A M. Aznan"A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset"ICSEC 2014 : The International Computer Science and Engineering Conference (ICSEC) 1-10.
[20] Anil K. Jain "Data clustering: 50 years beyond K- meansss "Pattern Recognition Letters (2010) 651–666.
[21] Mariscal, G., Marbán, Ó, & Fernández, C.(2010) A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137-166.
[22] S.B. Kotsiantis" Supervised Machine learning: A Review of classification Techniques", Emerging Artificial intelligence Applications in Computer Engineering I. Maglogiannis et al. IOS Press , 2007
[23] RuiXu ," Survey of Clustering Algorithms ", IEEE Transactions on Neural Networks Vol. 16, MAY 2005.
[24] Sherry Y. Chen and XiaohuiLiu.Journal of Information Science, © CILIP 2004 "The Contribution of Data Mining in Information Science".DRTC Workshop on Semantic Web 8th – 10th December, 2003 DRTC, Bangalore .
[25] Jun Lee, S. and Siau, K. (2001), "A review of data mining techniques", Industrial Management & Data Systems, Vol. 101 No. 1, pp. 41-46.
[26] Dietterich T.G. (2000) Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg
[27] Begum Çıgsar and DenizUnal , "Comparison of Data Mining Classification Algorithms Determining the Default Risk ",Hindawi Scientific Programming Volume 2019.
[28] Diseases Divya Sharma , Anand Sharma, Vibhakar Mansotra, A Literature Survey on Data Mining Techniques to Predict Lifestyle International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue VI, June 2017 IC Value: 45.98 ISSN: 2321-9653
[29] Comparative Analysis of Data Mining Tools and Techniques for Information Retrieval Amit Verma, Iqbal deep Kaur and Inderjeet Singh ISSN (Print) : 0974-6846 Indian Journal of Science and Technology, Vol 9(11).
[30] Sethunya R Joseph, Hlomani Keletso Letsholo," Data Mining Algorithms: An Overview" ISSN 2277-3061 Volume 15, International journal of computers and technology council for Innovative Research April, 2016.

[31] Abdulla H. Wahbeh et al. "A Comparison Study between Data Mining Tools over some Classification Methods". In *Proceedings of International Conference of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, pp. 18-26, 2011
[32] Rathee A, Mathur R P Survey on Decision Tree Classification algorithm for the Evaluation of Student Performance. International Journal of computers & Technology. 2013;
[33] Rohit Arora, Suman "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA"International Journal of Computer Applications, Volume 54, September 2012.
[34] XindongWu , Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang,Hiroshi Motoda , Geoffrey J. McLachlan ,Angus Ng ,Bing Liu, Philip S. Yu,Zhi-Hua Zhou , Michael Steinbach, David J. Hand, Dan Steinberg,"Top 10 algorithms in data mining " Springer-Verlag London Limited 2007.
[35] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998.
[36] Hlaudi Daniel masethe, Mosima anna masethe, "Prediction of heart disease using classification algorithms", proceedings of the world congress on engineering and computer science 2014 vol ii, 2014, san francisco, usa
[37] Shantakumar b. patil, "Extraction of significant patterns from heart disease warehouses for heart attack prediction", international journal of computer science and network security, vol.9, no. 2, february 2009
[38] Jayaram .Kkaregowda, A.G. ., Punya, v.,M.a., Manjunath, a.s., " Rule based classification for diabetic patients using cascaded K- meansss and decision tree c4.5.",International Journal of computer applications. 45–12, (2012)
[39] Vijayalakshmi, d., Thilagavathi, k., "An approach for prediction of diabetic disease by using b-colouring technique in clustering analysis", in: International Journal of applied mathematical research, 1 (4) pp. 520-530 science publishing corporation www.sciencepubco.com/index.php/ijamr (2012).
[40] Han, J., Rodriguze, J.C ., Beheshti, m., " Diabetes data analysis and prediction model discovery using rapidminer", Second International Conference on Future Generation Communication and Networking.96-9 (2008)
[41] Tackling the Poor Assumptions of Naive Bayes Text Classifiers Jason D. M. Rennie Lawrence Shih ,Jaime Teevanteevan, David R. Karger Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003
[43] Hua-Jun Zeng,Xuan-Hui Wang, Zheng Chen1 Hongjun Lu, Wei-Ying Ma" CBC:Clustering Based Text Classification Requiring Minimal Labeled Data Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03) 2003 IEEE
[44] Asha.T, S. Natarajan, and K.N.B. Murthy"A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification".
[45] Shekhar R. Gaddam, Vir V. Phoha, Kiran S. Balagani "K- meansss+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K- meansss Clustering and ID3 Decision Tree Learning Methods"IEEE Transactions on Knowledge & Data Enhineering, VOL. 19, NO. 3, March 2007.
[46] Mohammad Salim Ahmed ,Latifur Khan"SISC: A Text Classification Approach Using Semi Supervised Subspace Clustering"2009 IEEE International Conference on Data Mining Workshops.
[47] M.I. López, J.M Luna, C. Romero, S. Ventura "Classification via clustering for predicting final marks based on student participation in forums "Proceedings of the 5th International Conference on Educational Data Mining.
[48] Antonia Kyriakopoulou and Theodore Kalamboukis "Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems""ECML/PKDD 2008 Discovery Challenge.
[49] B.V. Sumana T. Santhanam" Prediction of diseases by Cascading Clustering and Classification "International Conference on Advances in Electronics, Computers and Communications (ICAECC) IEEE 2014
[50]. Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers.
[51]. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.
[52]. Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.
[53]. Chawla, N. V., Japkowicz, N., & Kolcz, A. (2003). Workshop learning from imbalanced data sets II. In *Proc. Int'l Conf. Machine Learning*.
[54]. Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, *48*(5), 1623-1637.
[55]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.
[56]. Zhou, L., & Lai, K. K. (2009). Benchmarking binary classification models on data sets with different degrees of imbalance. *Frontiers of Computer Science in China*, *3*(2), 205-216.
[57]. Sumana B. V., & Santhanam, T. (2016). Prediction of imbalanced data using Cluster based Approach *Asian journal of Information technology 15*(16):3022-3042, ISSN: 1682-3915 @ medwell journals

[58]. Chujai, P., Chomboon, K., Chaiyakhan, K., Kerdprasop, K., & Kerdprasop, N. (2017). A cluster based classification of imbalanced data with overlapping regions between classes. Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I, IMECS 2017, March 15 - 17, 2017, Hong Kong.
[59]. Elrahman, S. M. A., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*, 332-340.
[60]. Riquelme, J. C., Ruiz, R., Rodríguez, D., & Moreno, J. (2008). Finding defective modules from highly unbalanced datasets. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos*, *2*(1), 67-74.
[61]. Gupta, S., Parekh, B., & Jivani, A. (2019). A Hybrid Model of Clustering and Classification to Enhance the Performance of a Classifier. In *International Conference on Advanced Informatics for Computing Research* (pp. 383-396). Springer, Singapore.
[62]. Jin X., Han J. (2011). *K- means*ss Clustering.,Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.
[63]. Jakkula, V. (2006). Tutorial on Support Vector Machine (SVM). *School of EECS, Washington State University*, *37*.
[64]. Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS* International Transactions on Computer Science and Engineering, 30(1), 25-36.
[65]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of artificial intelligence research, 16, 321-357.
[66]. Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making, 5(4), 597-604.
[67]. Fernández, A., del Río, S., Chawla, N. V., & Herrera, F. (2017). An insight into imbalanced big data classification: outcomes and challenges. Complex & Intelligent Systems, 3(2), 105-120.
[68]. He, H., & Ma, Y. (Eds.). (2013). imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons.
[69]. García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining, In: Intelligent Systems Reference Library, 72, Springer, Berlin.
[70]. Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR), 49(2).
[71]. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463-484.
[72]. López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, *250*, 113-141.
[73]. Prati, R. C., Batista, G. E., & Silva, D. F. (2015). Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, *45*(1), 247-270.
[74]. Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92-122.
[75]. Zhang, H., & Li, M. (2014). RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, *20*, 99-116.
[76]. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, *School of Information and Computer Science.*
[77]. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2011). KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, *17*, 255-287.
[78]. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.
[79]. Demšar, J., Zupan, B., Leban, G., & Curk, T. (2004). Orange: From experimental machine learning to interactive data mining. In *European conference on principles of data mining and knowledge discovery* (pp. 537-539). Springer, Berlin, Heidelberg.
[80]. Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3), 5718-5727.
[81]. Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539-550.
[82] uci data set link
[83] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," Journal of Neurology, Neurosurgery and Psychiatry, vol. 79, no. 4, pp. 368–376, 2007.
[84] Little, M., and McSharry, P., Suitability of dysphonia measurements for tele-monitoring of Parkinson's disease. Nature Precedings. 1-27, 2008.

[85] Luukka, P., "Feature selection using fuzzy entropy measures with similarity classifier", Expert Systems with Applications, 38, 4600–4607, 2011.

[86] Li, D. C., Liu, C. W., Hu, S. C., "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets", Artificial Intelligence in Medicine, 52, 45–52, 2011.

[87] Farhan Mohammed , Xiangjian He , Jinjun Chen and Yiguang Lin, A Novel Model for Classification of Parkinson's Disease: Accurately Identifying Patients for Surgical Therapy Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

[88] Tsoulos IG, Mitsi G, Stavrakoudis A and Papapetropoulos S (2019) Application of Machine Learning in a Parkinson's Disease Digital Biomarker Dataset Using Neural Network Construction (NNC) Methodology Discriminates Patient Motor Status. Front. ICT 6:10. doi: 10.3389/fict.2019.00010.

[89] Duygu Kaya, "Optimization of SVM Parameters with Hybrid CS-PSO Algorithms for Parkinson's Disease in LabVIEW Environment," Parkinson's disease, vol. 2019, Article ID 2513053, 9 pages, 2019. https://doi.org/10.1155/2019/2513053.

7 PUBLICATIONS

1. Swaminarayan P**, Gupta S.** (2014)" Development of Geo-Visualized Information System for states of India based on rainfall using ArcGIS". International Journal on Recent and Innovation Trends in Computing and Communication, 2(7), 1894 - 1896.

2. **Gupta S**, Parekh.B, Undavia J, 2012, "A Fuzzy approach for Spam Mail Detection integrated with wordnet hypernyms key term extraction", International Journal of Engineering Research & Technology (IJERT) Volume 01, Issue 05 (July 2012).

3. **Gupta S**., Parekh B., Jivani A. (2019) "A Hybrid Model of Clustering and Classification to Enhance the Performance of a Classifier". In: Luhach A., Jat D., Hawari K., Gao XZ., Lingras P. (eds) Advanced Informatics for Computing Research. ICAICR 2019. Communications in Computer and Information Science, vol 1076. Springer, Singapore.
(**Scopus ,Web of Science)**

4. **Gupta S**. and Jivani, A. (2019). "A Cluster based Under-Sampling solution for handling imbalanced Data". International Journal on Emerging Technologies, 10(4): 160–170. **(Scopus and UGC-CARE List (No.212))**