International Journal on Emerging Technologies **10**(4): 160-170(2019)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

A Cluster Based Under-Sampling Solution for Handling Imbalanced Data

Subodhini Gupta[1] and Anjali Jivani[2]

[1]Assistant Professor, Department of Computer Application, Bhopal (Madhya Pradesh), India.
[2]Associate Professor, Department of Computer Science & Engineering, Baroda (Gujarat), India.

(Corresponding author: Subodhini Gupta)
(Received 02 September 2019, Revised 30 October 2019, Accepted 05 November 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The two indispensible task of data mining are clustering & classification. The integration of these tasks together can give better and accurate results compare to- unaccompanied. Taking the 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 problem in data mining. Most of the work is done on - between class imbalance problems. Very few researchers have addressed the problem of imbalanced distribution among data - within class. Our study is an approach to deal with these two problems (between- class imbalance distribution of data & within- class imbalance distributions of data) simultaneously. In this study we have proposed a cluster based sampling solution to classify the imbalanced data. We found that the proposed method is simple yet effective in order to classify imbalanced distribution of data.

Keywords: Between-class imbalanced, Classification, Clustering, Imbalanced Data, within class imbalanced, under sampling.

I. INTRODUCTION

"Data mining is a synonym to knowledge discovery in databases is a process of analyzing data from different perspectives and summarizing it into useful information" [1].

Classification and clustering are the two important tools, required to solve most of the data mining problems. Classification [2] is a two stage method where at first stage we built/trained model from historical training data sets with labeled class attributes. In the second stage we try to predict the class labels of new test datasets as accurately as possible. Fundamentally, it is a mapping from target function to attribute set of already labeled class.

Clustering similar to classification in which objects of data are grouped together without consulting a known class label [3]. 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 maximum intra-cluster similarity and minimum inter-cluster similarity.

Imbalance class distribution became noticeable with the application of data mining techniques in real-world applications. Chawla *et al.*, (2003) had grabbed attention for the first time in the workshop on their work "Learning from imbalanced datasets" [4]. This important issue is drawing the attention of the data mining community for decades and is identified as an open research problem. This paper presents a cluster based integrated framework (combines classification with clustering techniques) and its implementation that handles imbalanced datasets. Despite extensive research is going on handling imbalanced data, but most of the work is being done on balancing the imbalanced data, very insignificant work is being performed on between- class imbalances & within class [5] and Type-1 and 2 error [6].

The proposed framework gives a 3 fold solution to these problems.

1. Capable of handling the different degrees of imbalanced nature of data [7].
2. It balances imbalanced data using the under sampling method.
3. It is capable of handling between class imbalanced and within class imbalanced nature of data.

II. PRELIMINARIES AND BASIC DEFINITIONS

A. Imbalanced Data

Data classification problem is a challenging problem as it affects the performance of standard classifiers so drastically due to the unequal distribution of data among classes. A large number of examples belong to one class called Majority class and very few examples represent other class called Minority Class [8].

Fig. 1 depicts the imbalanced Distribution of instances in Majority and Minority. Red astrik symbols represent instances belonging to majority class and blue circles are instances belonging to minority class. It is clearly noticeable that majority class area is very dense compare to minority class area.

While evaluating the performance of any classifier the impact of imbalanced nature of real-world data cannot be ignored. Classifiers performance is always biased towards the majority class and considers minority class

instances as noise and do not give required weightage in the building of model.

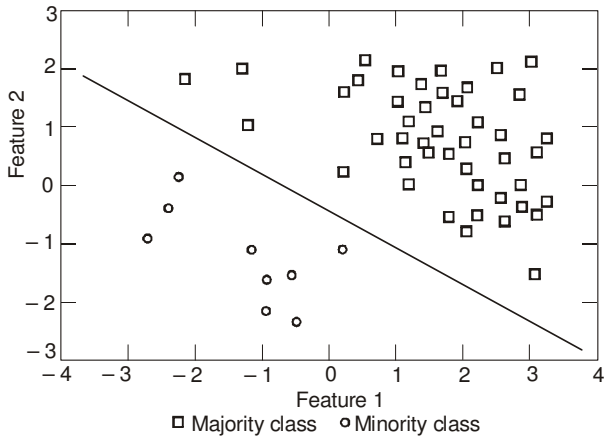

Fig. 1. Majority and Minority instances Distribution in Imbalanced class (Source: Chujai *et. al.*, [9]).

B. Between-class imbalance & within-class Imbalance

Between- class imbalance dataset [3] where the number of instances representing the Majority of class is extremely out-numbered the number of instances representing the Minority class;

Within - class imbalance dataset where a single class is composed of different sub-clusters and in these sub-clusters instances belonging to one sub-cluster are extremely outnumbered compare to other sub-clusters. Between-class imbalanced dataset problem exist in the two classes and within class imbalanced dataset is present in a single class.

C. Imbalance Ratio

Is the proportion of minority instance over majority instance [10].

$$\text{Imbalance Ratio} = \frac{\text{No. of instance in Majority class}}{\text{No. of instances in Minority class}}$$

In this study we have considered IR = 0.5 and 1.0. When we take IR = 0.5 it means sampled majority instances will be half of the minority instance i.e. if we have 100 instances from minority and 1000 from majority we will select only 50 instances from majority to make final training set. If IR = 1 means sampled majority instances will be equal to the number of minority instance *i.e.* if we have 100 instances from minority and 1000 from majority we will select only 100 instances from majority to make final training set.

D. False positive and False negative

False positive also referred to as TYPE-1 error and False Negative known as TYPE-2 errors are the undesirable outcomes of any classifier [11]. False positives take place when the classifier is predicting it as positive which is a false case; actually, it should be predicted as a negative case. On the other hand, when the classifier is predicting it as Negative, which is a false case; actually it should be predicted as positive case, a false positive outcome will result in unnecessary treatments – e.g: while considering a medical case study - a false negative will give a false diagnosis. The false positive outcome is very critical, where the disease is ignored can lead to the death of the patients because of no treatments.

E. Clustering

A clustering of D dataset is a partition of D into K clusters C1, C2, C3……Ck where i =1, 2, 3……..k and Ci ≠ NULL and D = Ci and Ci ∩ Cj =NULL i≠j i, j=1,2,3…….k.[12]

(1) K-Mean: K-mean method of data clustering is one of the oldest and yet very popular among Data Miners community. One reason for its popularity is; it is data driven, so less number of assumption are required. It uses greedy search strategy so it is able to divide large datasets into segments based on the number of cluster supplied as K [13]. It means full convergence of clusters. It aims to minimize squared error given by

Sum of Squared Error=$\sum_{i=1}^{k}\sum_{j=1}^{ik} \text{dist}(x_{i,}y_j)$

Distance will be calculated by Euclidean distance metric between x_i and y_j is given by

Euclidean distance=$\sqrt{\sum_{k=1}^{n}(x_{ik},y_{ik})^2}$

F. SVM Classifier

Support Vector Machine is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data [14]. In a high dimensional feature space Support Vector machines uses hypothesis space of a linear functions. We try to achieve a plane that has the maximum margin.

G. Methods of handling Imbalanced Data:

Fig. 2 displays two methods provided in the literature to tackle imbalanced class distribution.

(1) Data level approach: In this solution, data is modified to be applied on traditional classifiers.

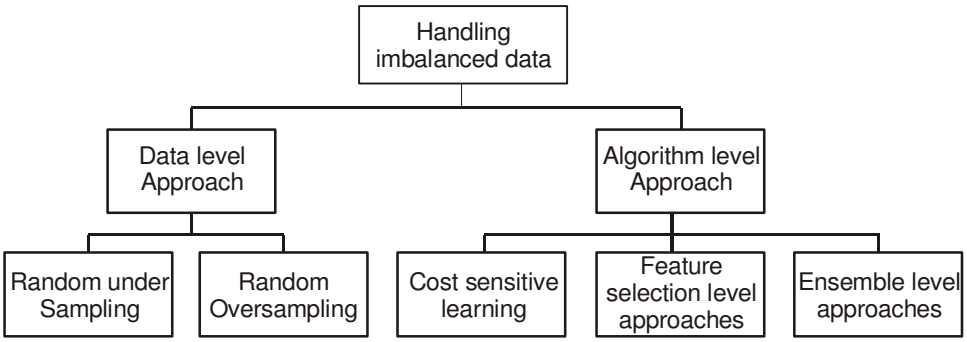

Fig. 2. Methods to handle Imbalanced Data.

(i) 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 [15]. Fig. 3 depicts how instances are randomly selected to increase or decrease the sample size.

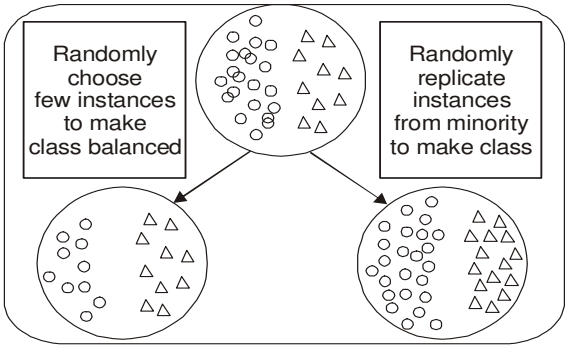

Fig. 3. Random Under-Sampling and Random Oversampling.

(ii) Synthetic Minority Oversampling Technique (SMOTE): Chawla *et al.*, (2002) [16] proposed a Synthetic Minority Oversampling Technique (SMOTE) - a remarkable research in the area of oversampling for classification of Imbalanced data is used in many applications. Feature based similarity is used to generate synthetic instances among minority instances. This method makes traditional classifier to enhance the decision boundary close to minority instances.

(2) Algorithm level Approach: Traditional classifiers are modified to deal with imbalanced data.

III. RELATED WORK

There are Rapid technological inventions in Data mining domain with an extra ordinary pace. Its implementation on real world problems on diverse areas - arise problem of imbalanced nature of data. It has been considered as one of the Top 10 challenging problems in data mining [17]. Many researchers have accepted the challenge and proposed their solutions.

These solutions basically fall in two categories: Data level and Algorithmic level [18].

A book which is in a form of paper collection edited by 'He and Ma' (2013) [19], It covers important issues such as sampling strategies, streaming data and active learning. A book by García *et al.*, [20] discussed about 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) [21]. A more specialized discussion thorough a survey on ensemble learning is given by Galar *et al.*, [22]. A global review on imbalanced data proposed by López *et al.*, (2013) [23] and an in-depth discussion on new perspectives of evaluating classifiers on Imbalanced datasets were presented [24]. A systematic review is done by Menardi and Torelli [25] on Class imbalance distribution. They have proposed re-sampling method that leads boosting and bagging to improve the accuracy for severe imbalanced data. Zhang and Li (2014) [26] performed experiments on three traditional classifier, used mean and standard deviation to generate samples for minority class. He stated that the oversampling influences the performances of traditional classifiers.

IV. EXPERIMENTAL INVESTIGATIONS

A. Datasets

A study for binary class distribution on 12 data sets openly available with different degrees of imbalance nature is conducted. The Table 1 contains the description of the data sets used for demonstrating the effectiveness of our proposed solution on various Parameters12 datasets were used from UCI or KEEL repository [27-28]. Number of instances, no. of attributes and degree of imbalanced distribution (imbalanced ratio) of the datasets are also given.

B. Experiment Setting

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

(1) WEKA: The WEKA workbench is a machine learning and data preprocessing tool, under 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 [29]. 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.

Table 1: Imbalanced Datasets with different degree 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

(2) Orange: Orange is a component-based data mining and machine learning software suite, it features a 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 [30].

C. Proposed Cluster Based Under-sampling

In Random Under-Sampling some instances are removed randomly [31]. 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. Here we are using clustering tool for sampling. 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 a 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 [32, 33] 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.

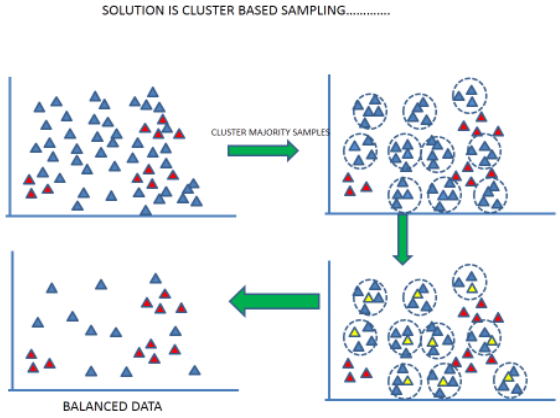

Fig. 4. Frame work of Cluster Based sampling.

D. Methodology

Overall process of transforming imbalanced data to balance can be divided into two phases: at first phase we will try to resolve Between –Class imbalance and Degrees of imbalanced distribution of data, for this we divide the whole dataset into two classes and will deal with them separately throughout the process. The first class called Majority class contains instance belongs to class containing large number of instances and the second class called Minority class belongs to a group containing less number of instances . In imbalanced distribution majority class size is always very large compare to minority. In second phase we will perform under sampling using clustering method here we will use K-means algorithm to partition majority class instances into K groups value of K will be optimized by silhouette plot. Now we will decide the imbalanced ratio for the resultant training set here in this experiment we have considered only two ratio they are 0.50 and 1.0 then we will calculate no of instances to be selected from each cluster using equation 1 after getting the number of instances we will randomly select required instances from each clusters.

In this process we will also take care for duplicity removal after this process we will have cluster representative instances from every cluster now we will merge all representative instances with minority class instances to make a Imbalanced training set we will repeat this process with IR = 1.

Then we will apply classifier and derive performance measuring parameters then we will compare these parameters with original dataset IR = 0.50 and IR = 1 to identify the better performances.

Dry run of algorithm with abalone dataset:

Dataset = Abalone

Total no. of instances = 4174

No. of Instances belonging to majority class=3813

No. of instances belonging to minority class=364

No of clusters optimized by silhouette plot K=2

No. of instances belonging to Cluster-1=2261

No. of instances belonging to Cluster-2=1552

No. of Instances to be sampled from cluster-1 is given by

$$SC_1^{inst} = IR \times \frac{MIN^{size}}{MAJ^{size}} \times NC^1$$

For IR=.50

$$SC_1^{inst} = 0.50 \times (\frac{364}{3813}) \times 2261$$
$$= 0.50 \times (0.09) \times 2261$$
$$= 0.50 \times 215$$
$$= 107$$

Total number of instances selected from cluster one for IR= 0.50 is 107

$$SC_2^{inst} = 0.50 \times (\frac{364}{3813}) \times 1552$$
$$= 0.50 \times (.09) \times 1552$$
$$= 0.50 \times 139$$
$$= 34$$

Total numbers of instances selected from cluster two for IR= 0.50 is 34.

For IR =1.0

$$SC_1^{inst} = 1.0 \times (\frac{364}{3813}) \times 2261 = 215$$

$$SC_1^{inst} = 1.0 \times (\frac{364}{3813}) \times 1552$$
$$= 139$$

Total number of instances selected from cluster-1 is 215.

Total number of instances selected from cluster-2 is 139.

Now we will merge these numbers of randomly selected instances from clusters to make balanced majority class having less number of instances representing overall population. To make a final training dataset we will merge these clusters with minority class instances.

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

Step 1: Segregate whole data set into $\mathrm{MIN}^{\mathrm{inst}}$and $\mathrm{MAJ}^{\mathrm{inst}}$
$\mathrm{MIN}^{\mathrm{size}}$ –No. of instance belongs to Minority class
$\mathrm{MAJ}^{\mathrm{size}}$ –No. of instances belongs to Majority instances.
In imbalanced datasets$\mathrm{MAJ}^{\mathrm{size}}$ >$\mathrm{MIN}^{\mathrm{size}}$

Step 2: Removing Outliers
$\mathrm{MIN}_{\mathrm{i}}^{\mathrm{inst}}$ where i=1, 2, 3, 4……..$\mathrm{MIN}^{\mathrm{size}}$
$\mathrm{MAJ}_{\mathrm{j}}^{\mathrm{inst}}$ where j=1, 2, 3, 4……..$\mathrm{MAJ}^{\mathrm{size}}$
If Distance ($\mathrm{MIN}_{\mathrm{i}}^{\mathrm{inst}}$,$\mathrm{MAJ}_{\mathrm{i}}^{\mathrm{inst}}$) =0
Remove $\mathrm{MAJ}_{\mathrm{j}}^{\mathrm{inst}}$ from $\mathrm{MAJ}^{\mathrm{size}}$ and update $\mathrm{MAJ}^{\mathrm{size}}$

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 silhouette plot to find most appropriate value
of K.
$\mathrm{MAJ}^{\mathrm{size}}$=$\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{C}_{\mathrm{i}}^{\mathrm{inst}}$

Step 5: NC^{i} no. of instances in Ith cluster
$\mathrm{SC}_{\mathrm{i}}^{\mathrm{inst}}$No. of instances to be sampled from each cluster is defined as:
$\mathrm{SC}_{\mathrm{i}}^{\mathrm{inst}}$ = IR×$\frac{\mathrm{MIN}^{\mathrm{size}}}{\mathrm{MAJ}^{\mathrm{size}}} \times \mathrm{NC}^{\mathrm{i}}$

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 given cluster
If Distance ($\mathrm{C}_1^{\mathrm{i}}$,$\mathrm{C}_{\mathrm{i}+1}^{\mathrm{i}}$) =0
Add $\mathrm{C}_1^{\mathrm{i}}$ in sampled training set and
Remove $\mathrm{C}_1^{\mathrm{i}}$ and duplicated $\mathrm{C}_{\mathrm{i}+1}^{\mathrm{i}}$ from cluster $\mathrm{C}_{\mathrm{i}}^{\mathrm{inst}}$
Repeat the process until we get the required instances from the cluster or instances of the cluster get exhausted.

Step 8: We will get k output clusters with selected no of quality instances. In order to get a balanced cluster we will merge
all output instances from each cluster with Minority instances to get a final Balanced training set.

Fig. 5 represents flow control of the proposed model. The classifier takes minority class instances as noise and does not consider them in building model so the classifier gets biased towards the majority class. Xiong (2010) stated that the Class Imbalance, class overlap added with high dimension data makes classifying task complicated and challenging [33-35].
As stated above that this model is Capable of giving 4 fold solutions. The first solution gives the capability of handling the different degree of imbalanced nature of data. In our approach, we divide the majority and minority instances into separate datasets and then deals majority class separately so in the first phase only we can handle the diverse.
Degree of imbalance distribution. It balanced imbalanced data using an under sampling method. Here in our approach, we deployed a cluster based selection to reduce the size of the majority class to make training set balanced to perform accurately on traditional classifiers. It is capable to handle between class imbalanced and within class imbalanced nature of data. Between class imbalance distributions is solved in the first phase when we classify majority and minority instances and within class imbalanced problem can be solved by making clusters from majority class and selecting uniform cluster representatives. 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.

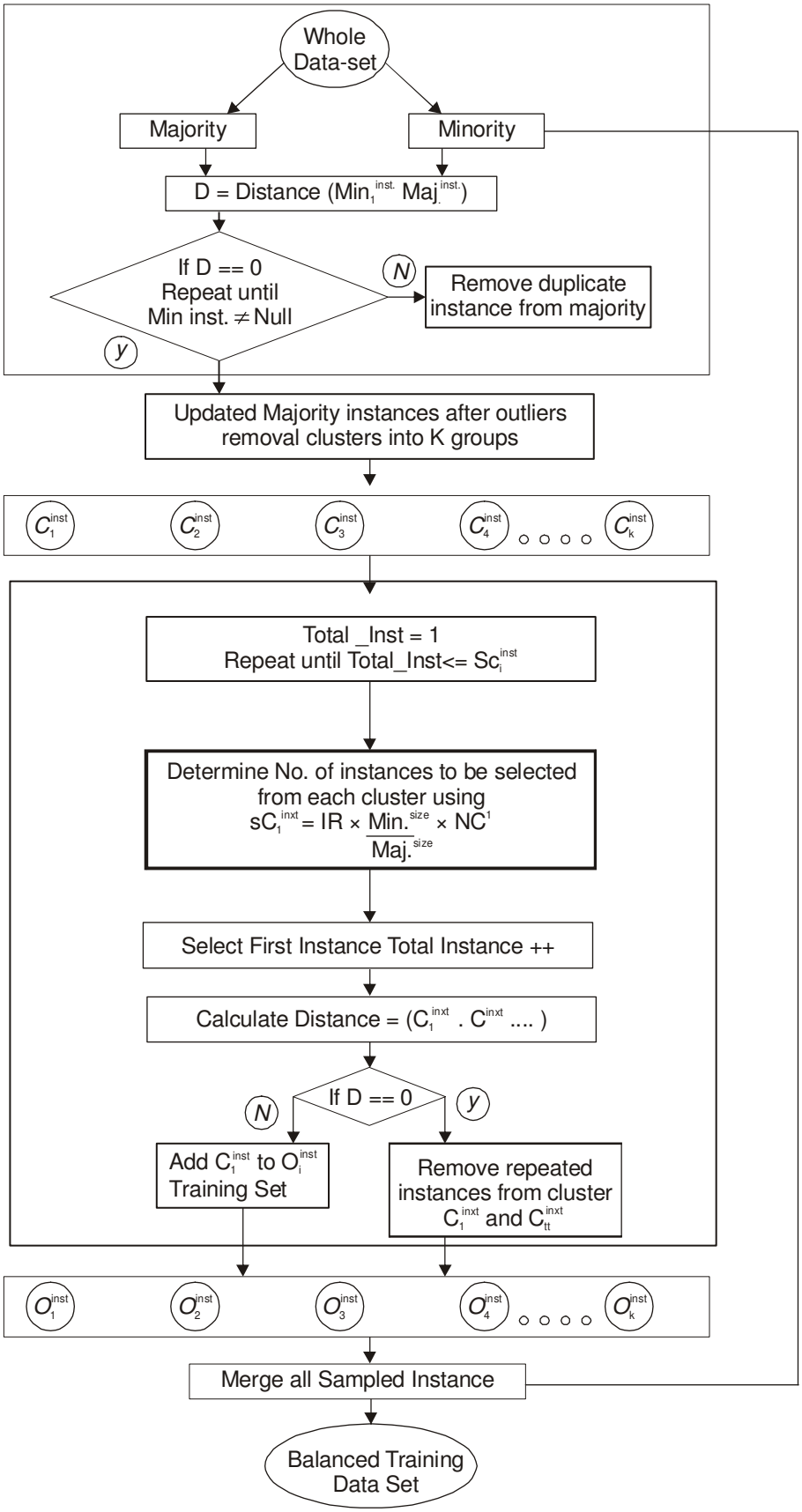

Fig. 5. Flow Chart for the proposed Algorithm.

Table 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	0.89	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 1	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

The Table 2 presents a comparative analysis of the performance of SVM classifier [36] on original imbalanced data of different degrees, Data with Imbalanced ration = 0.50 and data with Imbalanced ratio = 1. It also presents a collective information in order to identify which method's performance is better over other. Table 3 presents accuracy of SVM classifier for original dataset, with Imbalanced Ration of 0.50 and Imbalanced ration of 1 on 12 datasets. Fig. 6 is a pictorial representation of accuracy of SVM classifier for original dataset, with Imbalanced Ration of 0.50 and Imbalanced ration of 1 on 12 datasets. As we have discussed accuracy is not the perfect measure for imbalanced data.

In the graph we are getting high accuracy with few original imbalanced data sets but it does not prove overall good performance. Table 4 presents F-measure of SVM classifier for original dataset, with Imbalanced Ration of .50 and Imbalanced ration of 1 on 12 datasets. It can be easily identified that we are getting improved values for F-measure with proposed method. Fig. 7 is a pictorial representation of F-measure of SVM classifier for original dataset, with Imbalanced Ration of 0.50 and Imbalanced ration of 1 on 12 datasets. Table 5 presents FP rate of SVM classifier for original dataset, with Imbalanced Ration of 0.50 and Imbalanced ration of 1 on 12 datasets. It can be easily identified that we are getting improved values for FP rate with proposed method.

Table 3: Comparative Results for Accuracy measures for different Datasets with different imbalance Ratio.

S. No	Data set	Original data set	Balanced with IR = 0.50	Balanced with IR = 1.0
1.	Abalone	98	97	97
2.	Cleveland-0	95.9	80	96.1
3.	*E. coli*-3	89.5	92.3	81.1
4.	Glass-1	63.5	78	70
5.	Haberman	73	66.9	57
6.	New-Thyroid 1	92	90.1	98
7.	Page-Blocks 0	93	84	85
8.	Pima	77.3	77.3	71.7
9.	Wine Quality White	98.3	62	69.3
10.	Wisconsin	96.9	98.3	97
11.	Yeast 1	74.3	80	76
12.	Vowel	95.9	90	96

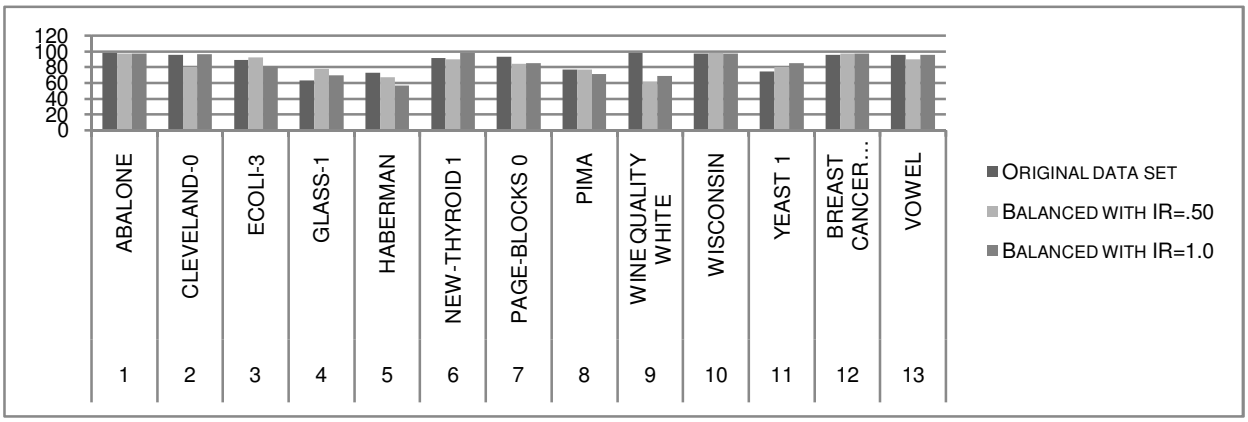

Fig. 6. Chart for accuracy.

Table 4: Comparative Results for F-Measures for different Datasets with different imbalance Ratio.

S.No	Data Set	Original data set	Balanced with IR = 0.50	Balanced with IR = 1.0
1.	Abalone	0.89	0.96	0.93
2.	Cleveland-0	0.97	0.84	0.98
3.	*E.coli*-3	0.94	0.94	0.83
4.	Glass-1	0.77	0.86	0.78
5.	Haberman	0.84	0.80	0.46
6.	New-Thyroid 1	0.95	0.93	0.98
7.	Page-Blocks 0	0.93	0.88	0.95
8.	Pima	0.62	0.84	0.70
9.	Wine Quality White	0.99	0.76	0.70
10.	Wisconsin	0.97	0.99	0.97
11.	Yeast 1	0.84	0.89	0.85
12.	Vowel	0.73	0.84	0.70

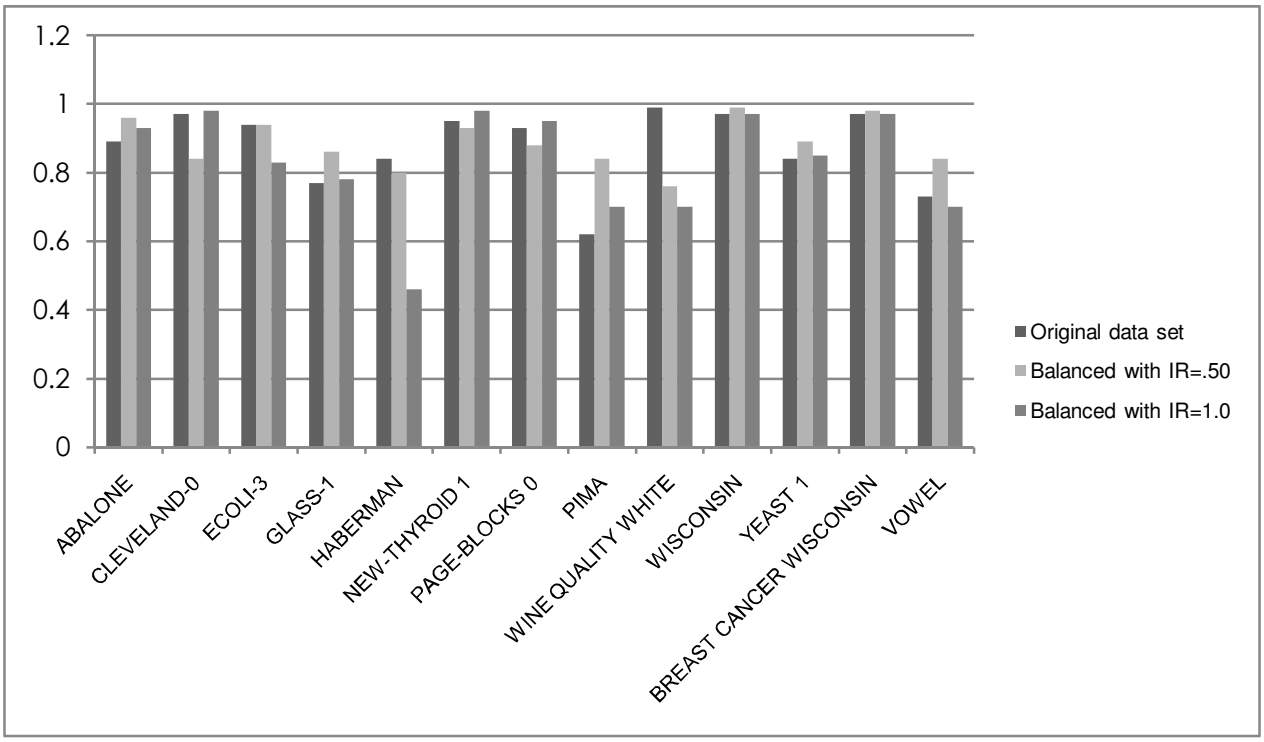

Fig. 7. Chart for F-Measure.

Table 5: Comparative Results for F-P rate for different Datasets with different imbalance ratio.

S. No	Data set	Original data set	Balanced With IR = 0.50	Balanced with IR = 1.0
1.	Abalone	0.16	0.07	0.058
2.	Cleveland-0	0.38	0.23	0.00
3.	*E. coli*-3	1	0.17	0.32
4.	Glass-1	1.0	0.59	0.54
5.	Haberman	1.0	1.0	0.2
6.	New-Thyroid 1	0.45	0.31	0.03
7.	Page-Blocks 0	0.59	0.18	0.081
8.	Pima	0.10	0.49	0.04
9.	Wine Quality White	1	1.0	0.33
10.	Wisconsin	0.15	0.02	0.03
11.	Yeast 1	0.81	0.23	0.31
12.	Vowel	0.50	0.11	0.23

Fig. 8 is a pictorial representation of F-P rate of SVM classifier for original dataset, with Imbalanced Ration of 0.50 and Imbalanced ration of 1 on 12 datasets. Table 6 presents precision of SVM classifier for original dataset, with Imbalanced Ration [37] of 0.50 and Imbalanced ration of 1 on 12 datasets. It can be easily identified that we are getting improved values for precision with proposed method.

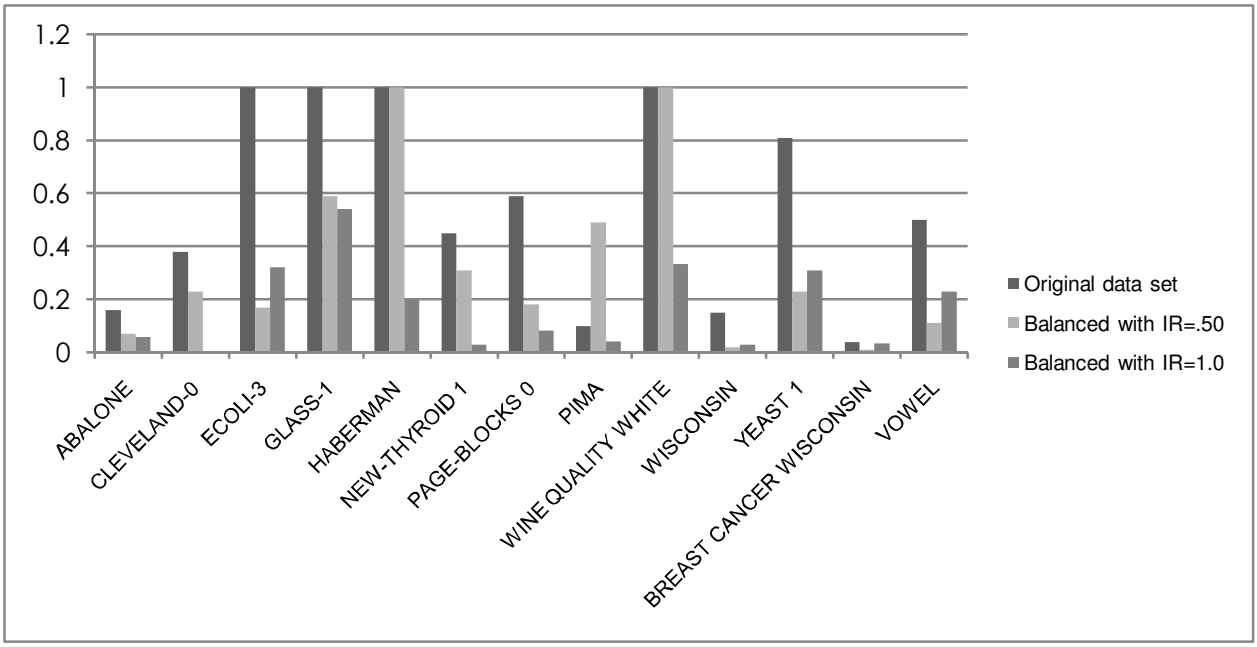

Fig. 8. Chart for F-P rate.

Table 6: Comparative Results for Precision rate for different Datasets with different imbalance ratio.

S.No	Data set	Original dataset	Balanced with IR = 0.50	Balanced with IR = 1.0
1.	Abalone	0.90	0.96	0.94
2.	Cleveland-0	0.9	0.84	1.0
3.	*E. coli*-3	0.89	0.91	0.75
4.	Glass-1	0.64	0.77	0.64
5.	Haberman	0.73	0.66	0.64
6.	New-Thyroid 1	0.91	0.87	0.97
7.	Page-Blocks 0	0.93	0.90	0.90
8.	Pima	0.74	0.78	0.73
9.	Wine Quality White	0.98	0.65	0.69
10.	Wisconsin	0.98	0.99	0.97
11.	Yeast 1	0.74	0.83	0.75
12.	Vowel	0.91	0.92	0.89

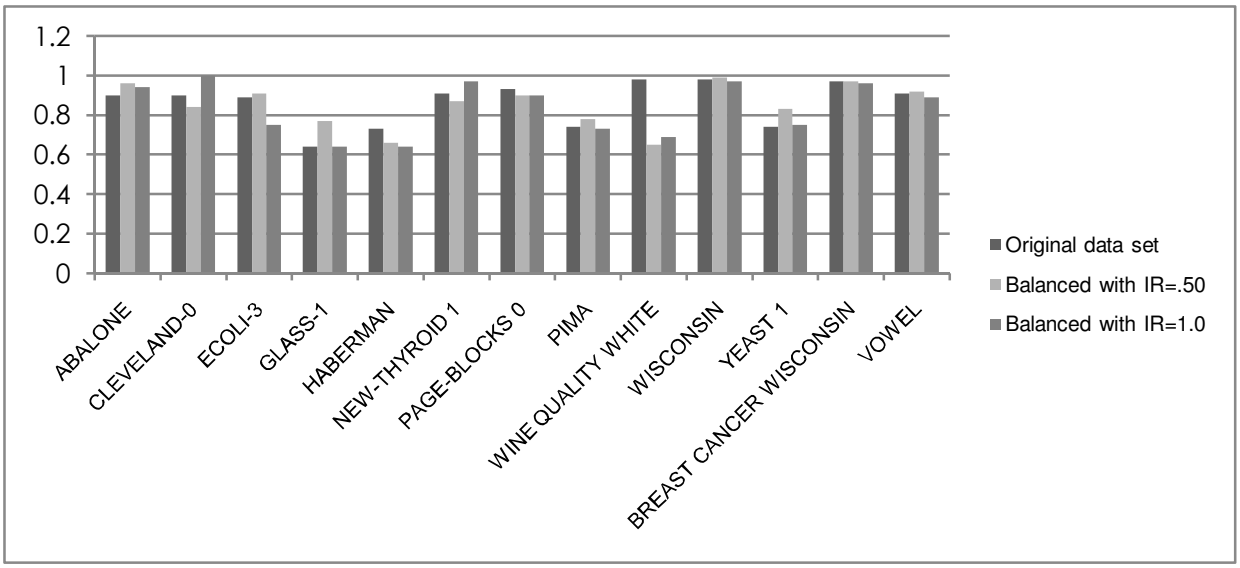

Fig. 9. Chart for precision.

Fig. 9 is a pictorial representation of Precision of SVM classifier for original dataset, with Imbalanced Ration of 0.50 and Imbalanced ration of 1 on 12 datasets. Table 7 presents ROC of SVM classifier for original dataset, with Imbalanced Ration of .50 and Imbalanced ration of 1 on 12 datasets. It can be easily identified that we are getting improved values for ROC with proposed method

Table 7: Comparative Results for ROC rate for different Datasets with different imbalance Ratio.

S. No	Data set	Original data set	Balanced with IR = 0.50	Balanced with IR = 1.0
1.	Abalone	0.90	0.98	0.97
2.	Cleveland-0	0.80	0.78	0.96
3.	*E. coli*-3	0.50	0.89	0.81
4.	Glass-1	0.49	0.74	0.70
5.	Haberman	0.50	0.50	0.57
6.	New-Thyroid 1	0.77	0.84	0.98
7.	Page-Blocks 0	0.70	0.83	0.85
8.	Pima	0.72	0.70	0.75
9.	Wine Quality White	0.50	0.46	0.69
10.	Wisconsin	0.96	0.78	0.77
11.	Yeast 1	0.57	0.80	0.68
12.	Vowel	0.80	0.93	0.74

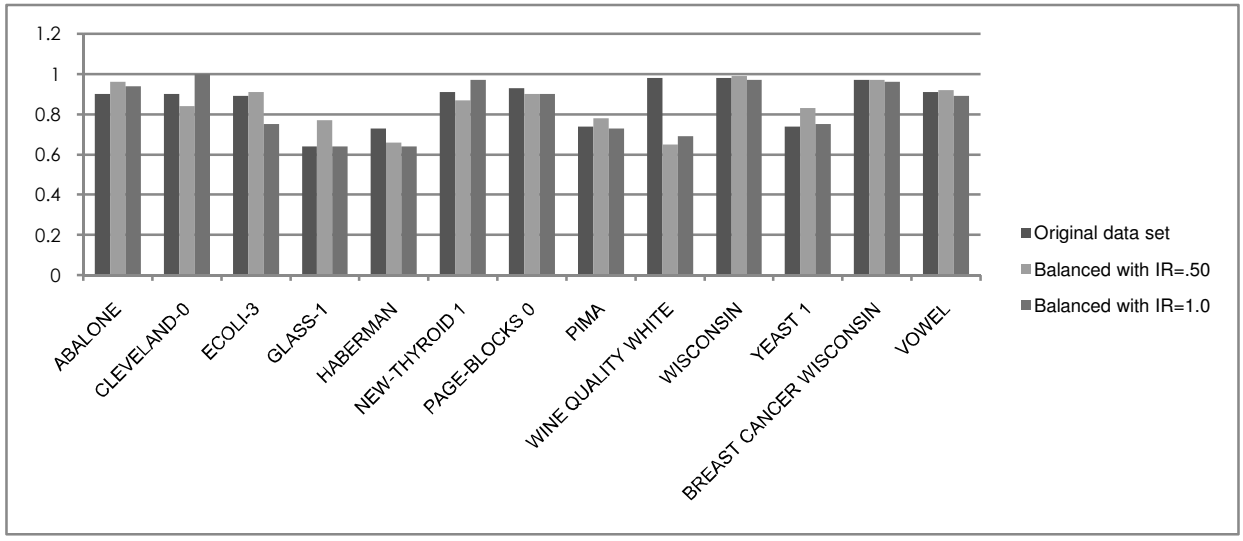

Fig. 10. Chart for ROC.

Fig. 10 is a pictorial representation of ROC of SVM classifier for original dataset, with Imbalanced Ration of 0.50 and Imbalanced ration of 1 on 12 datasets.

IV. CONCLUSION

In this research work cluster based under sampling method is applied on different degree of Imbalanced ratio over 12 data sets from UCI and KEEL repositories. Experimental results show the better performance of proposed algorithm on these data sets. It also deal 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.

V. FUTURE SCOPE

These experiments are conducted for Binary classification and with only SVM classifier. Other classifiers can also be considered with multi class classifications and False positive and False negative errors.

REFERENCES

[1]. Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers.

[2]. 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.

[3]. Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.

[4]. Chawla, N. V., Japkowicz, N., & Kolcz, A. (2003). Workshop learning from imbalanced data sets II. In *Proc. Int'l Conf. Machine Learning*.

[5]. 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.

[6]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

[7]. 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.
[8]. 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
[9]. Chujai, P., Chomboon, K., Chaiyakhan, K., Kerdprasop, K., & Kerdprasop, N. (2017). A cluster based classification of imbalanced data with overlapping regions between classes. In *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, *1*.
[10]. Elrahman, S. M. A., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*, 332-340.
[11]. 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.
[12]. 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.
[13]. Jin X., Han J. (2011). *K*-Means Clustering.,Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.
[14]. Jakkula, V. (2006). Tutorial on Support Vector Machine (SVM). *School of EECS, Washington State University*, *37*.
[15]. Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), 25-36.
[16]. 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.
[17]. Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5*(4), 597-604.
[18]. 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.
[19]. He, H., & Ma, Y. (Eds.). (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.
[20]. García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining, In: *Intelligent Systems Reference Library*, *72*, Springer, Berlin.
[21]. Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, *49*(2).1-48.
[22]. 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.
[23]. 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.
[24]. 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.
[25]. Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92-122.
[26]. Zhang, H., & Li, M. (2014). RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, *20*, 99-116.
[27]. 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.*
[28]. 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.
[29]. 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.
[30]. 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.
[31]. Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3), 5718-5727.
[32]. 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.
[33]. Rout, N., Mishra, D., & Mallick, M. K. (2018). Handling imbalanced data: a survey. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications* (pp. 431-443). Springer, Singapore.
[34]. Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: a review. *Int. J. Advance Soft Computer Appl.*, *7*(3), 176-204.
[35]. Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence* (IC-A1'2000): Special Track on Inductive Learning, Las Vegas, Nevada.
[36]. Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing & Management*, *47*(4), 617-631.
[37]. Maimon, O., Rokach, L., & Chawla, N. (2005). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 853-867.

How to cite this article: 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.