Divide and Conquer Solution to Class-Imbalance Problem in Classification pattern

Ms. Ashwini Jigalmadi

Abstract— Large dataset and class imbalanced distribution of samples across the data classes are intrinsic propertied of the problems to be faced in the applications like bioinformatics, network security and text mining. The class imbalanced problem appears in the dataset, classification categories are not represented with approximately equal number of instances. In this paper, we have explored the solution to the problem of imbalanced representations of the classes in the dataset. In this method, instance selection is applied concurrently to the small class-balanced subsets of the training data. Then, subsets are combined based on the voting score calculated from the optimized pair of thresholds of minority and majority classes. We used support vector machine (SVM) and kNN classifier to perform the experiments on the dataset for analyzing the performance of proposed algorithm. On comparison, it is observe that proposed algorithm outperform the random sampling method. Further, proposed algorithm has linear computational complexity and can be easily be implemented using parallelism to have real-time performance.

Index terms— Divide and conquer, Imbalance-class problem, Instance sampling, kNN classifiers, Machine learning, Majority, Minority, Voting.

1 INTRODUCTION

With the continuous expansion of data availability in various networked, complex, and large-scale systems, such as Internet, security, surveillance, and finance, it becomes critical to advance the fundamental understanding of knowledge discovery and analysis from raw data to support decision-making processes [1]. Although many methods have been proposed for dealing with class-imbalance data sets, most of these methods are not scalable to the very large data sets common to those research fields.

The class imbalance problem is one of the (relatively) new problems that emerged when machine learning matured from an embryonic science to an applied technology, amply used in the worlds of business, industry and scientific research. Most classification methods suffer from an imbalanced distribution of training instances among classes and most learning algorithms expect an approximately even distribution of instances among the different classes and suffer, to different degrees, when that is not the case. Dealing with the class-imbalance problem is a difficult but relevant task as many of the most interesting and challenging real-world problems have a very uneven class distribution. The solution to these kind of problems are achieved either by modifying the learning algorithm, where cost is biased towards the one of the class, or by manipulating the training data sets, where resampling is applied, or by combining both. However, there exist a main advantage using the solutions applied at training data. The summary of solutions are depicted in figure 1.

For the similar problem, we explore a new framework called oligarchic instance selection, which is specifically designed for class imbalanced data sets. One of the distinctive features of many common problems in data mining applications is the uneven distribution of the instances of the different classes. In extremely active research areas, such as artificial intelligence in medicine, bioinformatics, or intrusion detection, two classes are usually involved: a class of interest or a positive class, and a negative class that is overrepresented in the data sets. This is usually referred to as the class-imbalance problem.

The method has two major objectives:

- Improving the performance of previous approaches based of instances selection for class-imbalanced data sets.
- Developing a method that is able to scale up to very large, and even huge, problems.

This project aims at developing method that is both scalable and able to sample the most relevant instances to deal with class-imbalanced data sets. Scalability will be achieved using a divide-and-conquer approach. The ability to sample instances to deal with class-imbalanced data sets will be achieved by means of the combination of several rounds of instance selection in balanced subsets of the whole data set.

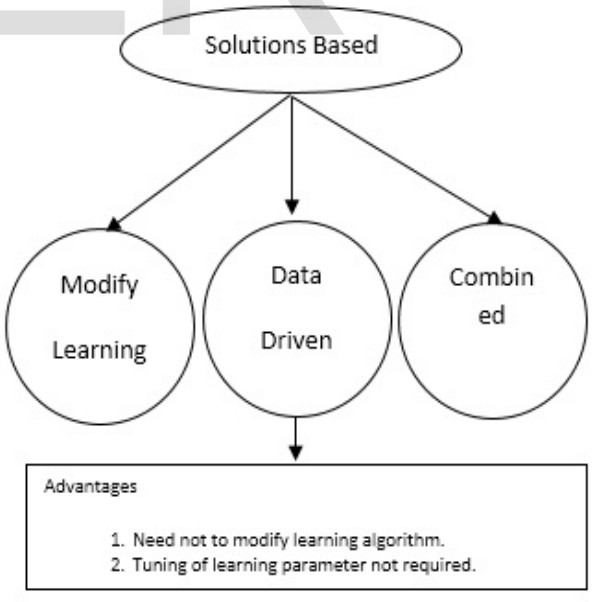

Figure 1: Types of solutions for class-imbalance problem.

The remaining part of the paper is organized as follows. Next section II presents the related work in back ground. The problem stamen is briefly introduced in the section III. The methodology supported by mathematical model and set theory and snippet of algorithm is depicted in the section IV. The

implementation details is described in the section V. Section VI discusses the results and graphs. Finally, paper is concluded by highlighting the main observations and giving future direction for research work.

2 REMAINING CONTENTS

2.1 Related Work and Background

The data driven methods has advantage of not modifying the algorithm of classifier learning. This also saves the effort of tuning the various parameters of learning algorithm. In general data driven methods of solving the problem of class-imbalance applies under sampling to the majority class or oversamples the minority class or does by combining both. The process of oversampling or under sampling of instances can be done using the random sampling or by searching the least or most useful instances from training dataset.

In [2], it is proven that under sampling the majority class gives better results than oversampling the minority performed using sampling with replacement. However, combining under sampling of the majority class with oversampling the minority class instances does not yield better performance compared to the under sampling of the majority class alone. This is shown in [3] and concluded that it is happened because oversampling does not add any new information of the type of inputs to the classifier. In [4] and [5], authors have proven sampling as a very efficient method dealing with class-imbalanced datasets. In one-sided selection (OSS), instances from majority class are moved and this technique is applied in [6]. However, in this method as there is no sampling involved in the minority class instances, it doesn't have capability to remove the malfunction to be caused by harmful sample from minority class. In [7], the sampling of instances is optimized using evolutionary computations. However, evolutionary computation may become very expensive in computation for large and very large datasets. The scalability which is very important [8] in large dataset problem becomes near-impossible in evolutionary techniques [9]. The instance selection in sampling the dataset is achieved using voting score in [10, 11]. Various instance selection methods to achieve the balanced dataset from imbalanced-class training data are presented in [12, 13, 14].

2.2 Problem Statement

The data driven methods has advantage of not modifying the algorithm of classifier learning. This also saves the effort of tuning the various parameters of learning algorithm. In general data driven methods of solving the problem of class-imbalance applies under sampling to the majority class or oversamples the minority class or does by combining both. The process of oversampling or under sampling of instances can be done using the random sampling or by searching the least or most useful instances from training dataset.

The recognition or classification of classes is a two-steps problem, namely, feature extraction or data representation and classifying step. Once the every sample is represented by variable, it is given to the classifier, whose outputs the label of recognized class. Before, classification of unknown sample, classifier is trained with training data. This is also called as a process of machine learning. The training data samples are expected to be evenly distributed across all the classes. However, there is a vast amount of data available in some sources like internet web, social networking, blogs, health care search etc. Due to unstructured way of generation of data, there is a high possibility of having data samples un-evenly associated across the classes.

In particular, there is an additional requirement of scalability in the sense that the balancing the class-imbalance problem should be applicable to large dataset. This implies the two requirements to be handled in the applications, where imbalanced-class problem dominates.

1. Selecting the instances from the classes such that imbalanced training data can be reduced to the class-balance data.
2. The transforming imbalanced training data into balanced class data need to be consistent towards large dataset

2.3 Methodology

We have used two methods for the evaluation of the OligoIS:

- Selection of samples according to the voting
- Selection of Sample according to the Euclidean Distance

For the comparison we have used following two methods:

- Random Under sampling for the with Balanced Dataset
- Random Under sampling with imbalanced Dataset

2.3.1 Mathematical Model and Set Theory:tle and authors

Norrmaly, there will be two types of data in the training dataset. One is minority classes which are underrepresented in the sense the number of instances associated with minority classes will be very less. On the other hand majority class will be over represented, with the ratio of 1:1000 and even sometimes 1:10,000.

The training dataset is partitioned into small subsets and decomposition process is given by and is shown in figure 2a.

$$T = \bigcup_{j=1}^{t} Di$$

This is achieved by using random sampling. Each subset is balanced by adding randomly selected instances of the minority class. To include the enough minority instances in each of the subset, the size of each subset satisfies, $S \leq 2n^{+}$.

Instance selection in each subset is done and votes for selected instance is recorded. The vote of instance is defined as the number of times particular instance is selected. Once, this process is finished, the instances with majority votes are kept. The threshold for votes is calculated as follows.

For threshold t, selected instance S(t), then

$$f(S(t)) = 2r(S(t)) + (1-\alpha)\,a(S(t))$$

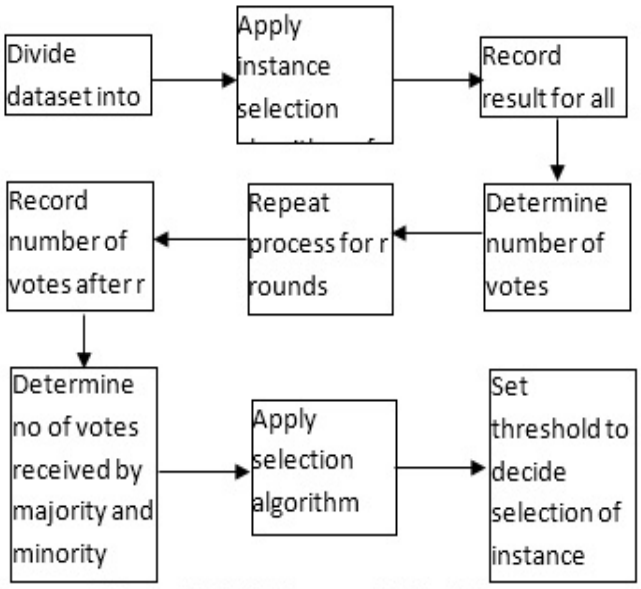

Figure 3: Block Diagram of Oligo Process

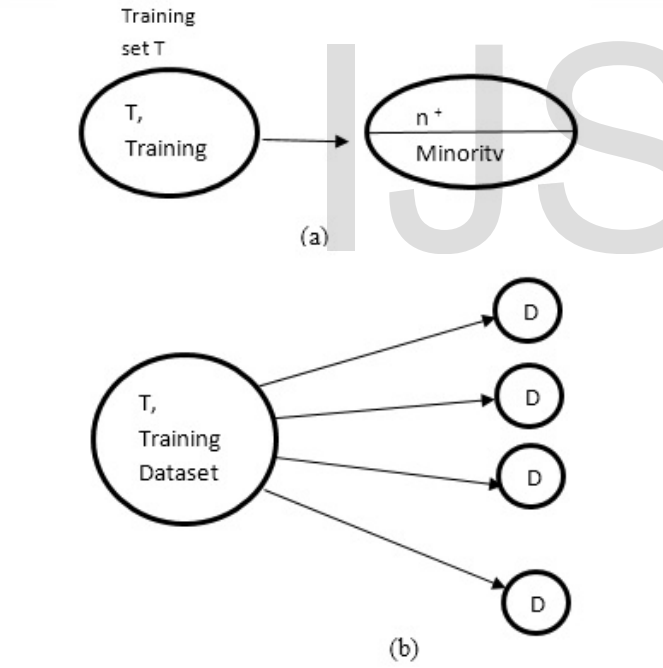

Figure 2: a) Training Dataset Sets b) Training dataset decomposition

Where $r(S(t))$ is the reduction achieved with threshold to select $(S(t))$.

$a(S(t))$ is the accuracy achieved with the instance in $(S(t))$ using SVM classification.

To account for class imbalanced in subset, above formulation is modified as below:

Two thresholds are t+ and t- for minority and majority respectively.

Thus, we have two equations

$$f(S(t^+)) = \alpha r\,(S(t^+)) + (1-\alpha)(S(t^+))$$
$$f(S(t^-)) = \alpha r\,(S(t^-)) + (1-\alpha)(S(t^-))$$

Combining this we get equation for pair of thresholds.

$$f(S(t^+.t^-)) = \alpha r\,(S(t^+.t^-)) + (1-\alpha)(S(t^+.t^-))$$

2.3.2 Methods for the Evaluation

2.3.2.1 OligoIS with Voting

In this method we have used random selection in each of the subset for many number of rounds. Samples which has got more number of votes are selected. This process outputs a final dataset which both majority and minority are present in equal numbers. Algorithm for this method is shown below.

2.3.2.2 OligoIS with ED based Selection

In this method we have used random selection in each of the subset for many number of rounds. Samples which has got more ED from the other class are selected. This process gives a final dataset which both majority and minority are present in equal numbers.

Algorithm for that is as follows:

Data: A training set T = {(x1, y1), . . . , (xn, yn)}, subset

size s, and number of rounds r.

Result: The set of selected instances S ⊂ T.

for i = 1 to r do

1. Divide instances into ns disjoint subsets

Di :Ui Di = T of size s

for j = 1 to ns do

2. Apply instance selection algorithm to Dj

3. Store the Euclidean Distances of samples from other class Dj

end

end

4 Obtain thresholds of ED to keep an instance from the

minority, t+, and the majority, t−, classes

5 S={xi ∈ T | (ED(xi) ≥ t+ and xi ∈ C+) or (ED(xi) ≥ t− and xi ∈ C−)}

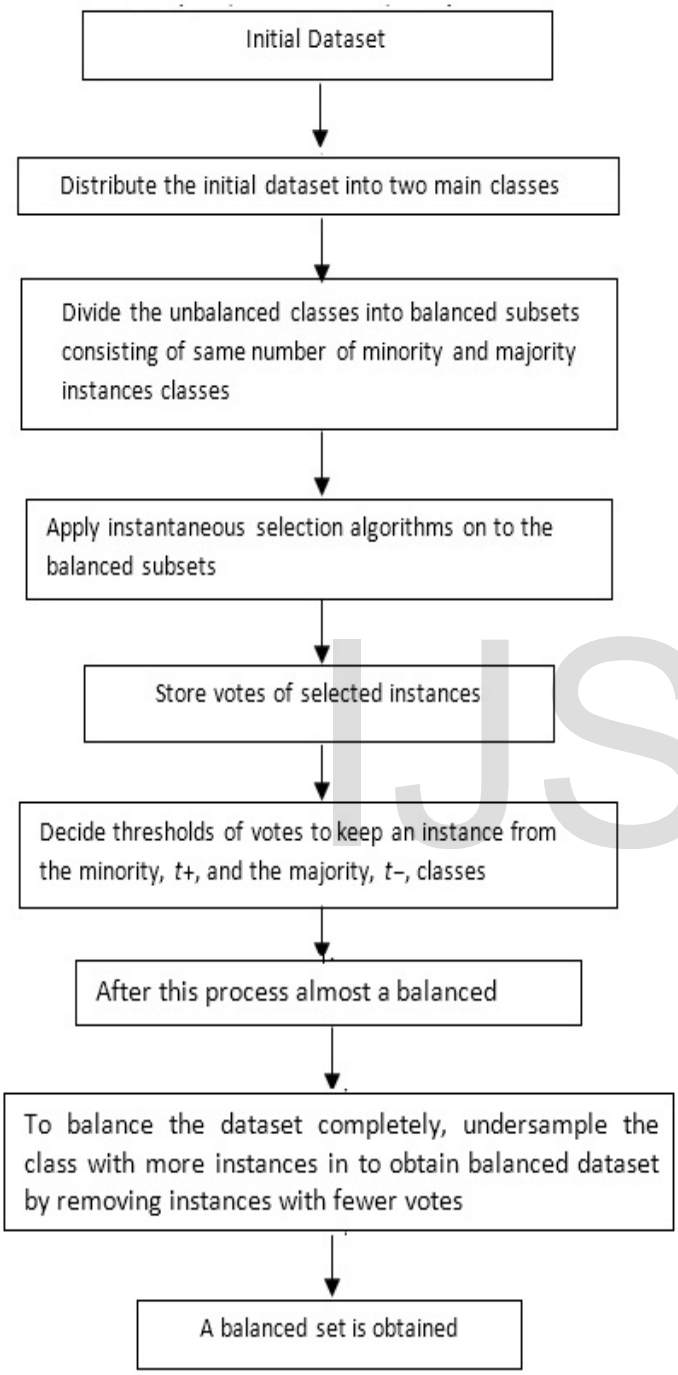

Figure 4: Flowchart for the OligoIS with Voting

6. Under sample the class with more instances in S to obtain Sbalanced removing instances with fewer votes

if f(Sbalanced) ≥ f(S) then

S = Sbalanced

end

7. return S

2.3.2.3 Random Selection with balanced dataset

In this method we have used random selection without any subset mechanism. Then the under sampling is done on the samples. Final subset consist of randomly selected samples with equal no of majority and minority samples.

Algorithm for that is as follows:

Data: A training set T = {(x1, y1), . . . , (xn, yn)}, subset size s, and number of rounds r.

Result: The set of selected instances S ⊂ T.

1. Apply instance selection algorithm to Tj
2. Select the Equal Number of instances from both classes
3. S=Random(T)
4. if f(Sbalanced) ≥ f(S) then

S = Sbalanced

5. return S

2.3.2.4 Random Selection with balanced dataset

In this method we have selected the dataset using random selection result dataset is the imbalanced dataset.

Algorithm for that is as follows:

Data: A training set T = {(x1, y1), . . . , (xn, yn)}, subset size s, and number of rounds r.

Result: The set of selected instances S ⊂ T.

1. Apply instance selection algorithm to Tj
2. Randomly select instances from both classes not necessary to be in equal amount
3. S=Random(T)
4. return S.

3 Experimental Results

We have used the various datasets obtained from UCI Machine Learning Repository [14].In order to align the dataset to two class imbalanced problem we have selected the dataset where samples of two classes are present. We performed the pattern recognition experiment on this eight datasets. The name of the datasets are adult, German, Haberman, hepatitis, magic04, ozone1hr, ozone8hr, Pima. The specification of each

of these datasets in terms of no of attributes, no of classes, no of samples and Imbalance Ratio (IR) is depicted in table.

The class-imbalanced problem is mainly due the fact that in real life applications based on binary-class recognition will have uneven distribution of samples between the two classes. This will deteriorate the performance of recognition system. To overcome this problem we have implemented the methods as discussed in earlier section. These methods are 1)OligoIS with selection of samples according to the voting 2)OligoIS with selection of Sample according to the Euclidean Distance 3)Random Under sampling for the with Balanced Dataset 4)Random Under sampling with imbalanced Dataset. We also propose the method of SVM based classification and implemented with different types of instance selection methods. The recognition accuracy for each of this methods across all the datasets are presented here. The proposed method is compared with the kNN based classification. The results with recognition accuracy obtained with KNN classification with different methods of instance selection across all the datasets are shown in Table 1. The results with recognition accuracy obtained with SVM classification with different methods of instance selection across all the datasets are shown in Table 4.

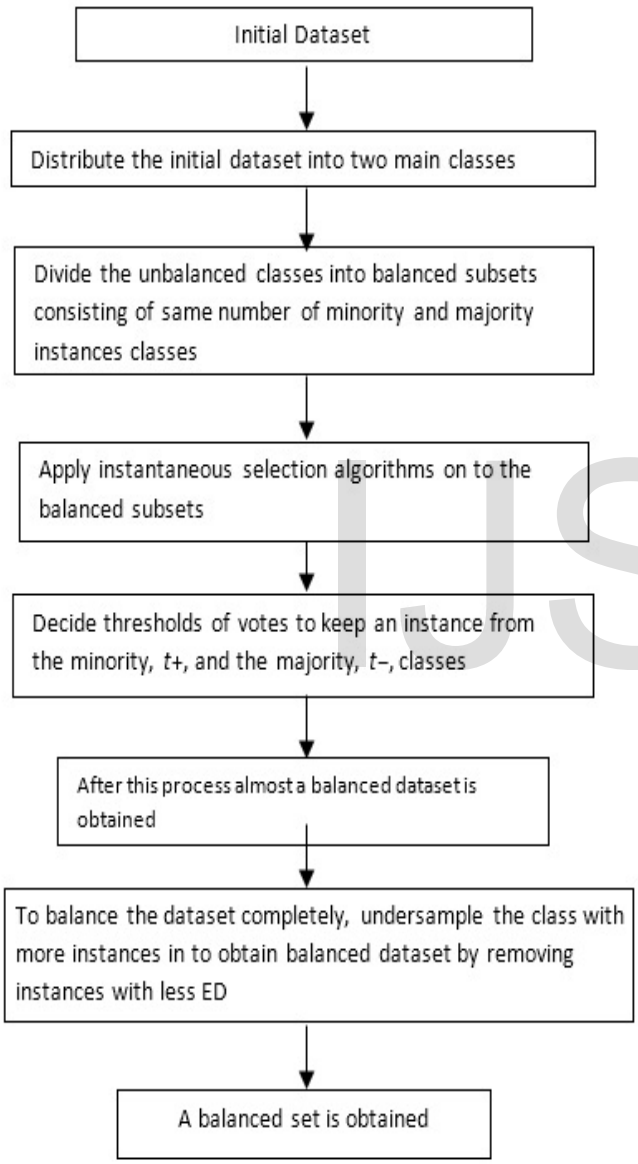

Figure 5: Flowchart for OligoIS with ED

For each of the dataset used in all this experiments have considered the 80% of the all the samples available for the training purpose the remaining 20% samples are used for the testing algorithm. In order to compare results visually we have also plotted the graphs for kNN and SVM with different instance section methods across all the datasets are plotted in figure 6 and 7 respectively.

We have also found the error rate deviation in the Oligo is far less that the other instance selection algorithms. As shown in figure 8. By observing the results we can say that OligoIS outperforms the various instance selection algorithms.

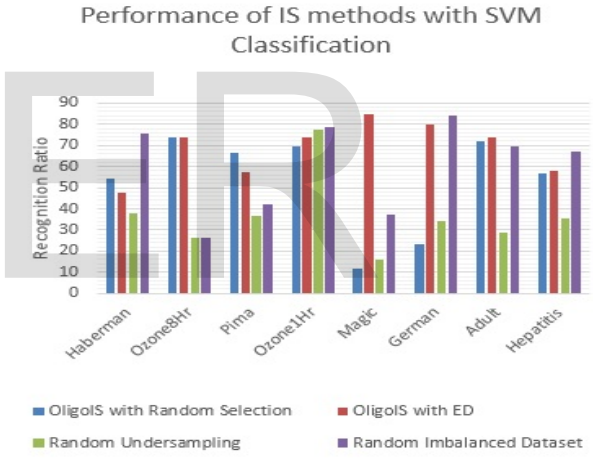

Figure 6. Performance of IS methods with SVM Classification

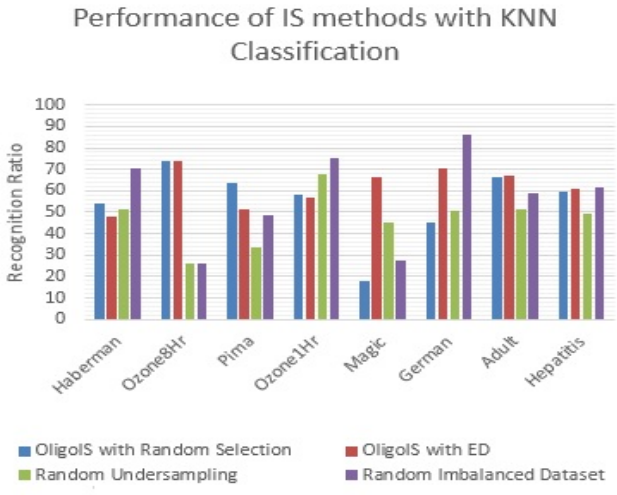

Figure 7. Performance of IS methods with SVM Classification

Table 1: Information about Datasets used

Dataset	No of Samples	No of Attributes	Name of Classes	IR
Haber-man	306	3	Survived more than 5 years or not	1:23
Ozone8Hr	2534	72	Ozone day or Not	1:15
Pima	768	8	Patient Having Diabetics or not	1:2
Ozone1Hr	2536	72	Ozone day or Not	1:34
Magic	19020	10	Gamma or Hadron	1:2
German	1000	20	Person is capable of returning money or not	1:3
Adult	48882	14	Income of more than 50k or not	1:4
Hepatitis	155	19	Die or Live	1:4

Dataset Name	Dataset Specifications				OligoIS with Random Selection	OligoIS with ED	Random Under sampling	Random Selection with Imbalanced Dataset	Imbalanced Dataset
	# +ve Sample	# -ve Samples	# Selected +ve samples	# Selected -ve samples					
Ozone8Hr	161	2375	10	10	73.77049	73.77049	26.22951	26.22951	68.85246
Adult	11687	37155	4630	4630	66.02564	67.30769	51.28205	58.97436	66.02564
Pima	269	501	20	20	63.63636	51.51515	33.33333	48.48485	63.63636
Hepatitis	71	86	27	27	59.5	61	49.5	61.5	62.5
Ozone1Hr	74	2464	12	12	58.4866	57.01524	67.70888	74.93431	77.16763
Haberman	81	225	39	39	54.30256	47.85947	51.28678	70.16768	72.85793
German	300	700	106	106	44.99018	70.72692	50.49116	86.44401	91.3556
Magic	6689	12333	3386	3386	18.23529	66.66667	45.4902	27.64706	95.29412

Table 2: Performance of IS methods with KNN Classification

Dataset Name	Dataset Specifications				OligoIS with Random Selection	OligoIS with ED	Random Under sampling	Imbalanced Dataset	Imbalanced Dataset
	# +ve Sample	# -ve Samples	# Selected +ve samples	# Selected -ve samples					
Ozone8Hr	161	2375	10	10	73.77049	73.77049	26.22951	26.22951	73.77049
Adult	11687	37155	4630	4630	71.79487	73.71795	28.84615	69.23077	73.71795
Ozone1Hr	74	2464	12	12	69.23279	73.69942	77.37782	78.61272	78.98056
Pima	269	501	20	20	66.66667	57.57576	36.36364	42.42424	60.60606
Hepatitis	71	86	27	27	57	58	35.5	67	73.5
Haberman	81	225	39	39	54.26571	47.85947	37.78023	75.48676	75.96585
German	300	700	106	106	22.98625	79.96071	34.18468	84.28291	80.74656
Magic	6689	12333	3386	3386	11.56863	84.70588	15.88235	37.2549	88.23529

Table 3: Performance of IS methods with SVM Classification

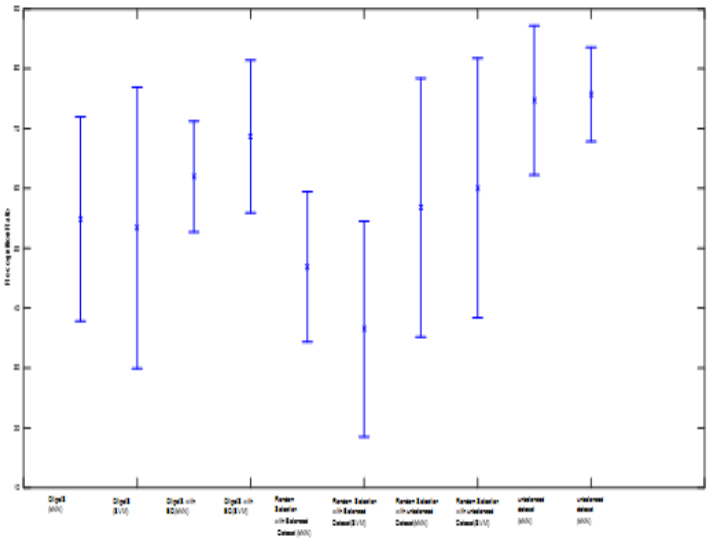

Figure 8: Error Graph for instance selection methods

4 Discussion

From results we can see that the OligoIS with the Euclidean distance gives more stable results that OligoIS. This makes it more suitable for creating the dataset in various domains.

5 Conclusion

The class imbalance problem is one of the (relatively) new problems that emerged when machine learning matured from an embryonic science to an applied technology, amply used in the worlds of business, industry and scientific research. In this paper, we have explored the solution to the problem of imbalanced representations of the classes in the dataset. In this method, instance selection is applied concurrently to the small class-balanced subsets of the training data. Then, subsets are combined based on the voting score calculated from the optimized pair of thresholds of minority and majority classes. We used support vector machine (SVM) and kNN classifier to perform the experiments on the dataset for analyzing the performance of proposed algorithm. On comparison, with other four methods it is observe that proposed algorithm outperform the random sampling method. Further, proposed algorithm has linear computational complexity and can be easily be implemented using parallelism to have real-time performance. The future work could be carried out in the direction of implementing the solution of imbalanced-class problem in parallel and analyzing the methods which can give real time performance in large dataset

Acknowledgment

The authors would like to thank the Department of Computer Engineering of JSPM's Rajarshi Shahu College of Engineering, as well as researchers for making their resources available and teachers for their guidance. We are thankful to the authorities Board of Studies Computer Engineering of Savitribai Phule Pune University. We are also thankful to reviewer for their valuable suggestions. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

- *Ashwini Jigalmadi is currently pursuing masters degree program in compueter science engineering inSavitribai Phule Pune University, India, PH-02064739803. E-mail: ashwini.bedadurge@gmail.com*
- *Dr.P.K.Deshmukh , India, PH-02064739803. E-mail:pkdeshmukh9e@gmail.com*

References

[1] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[2] N. V. Chawla, W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, Jan. 2002.
[3] C. Ling and G. Li, "Data mining for direct marketing problems and solutions," in Proc. 4th Int. Conf. KDD, New York, 1998, pp. 73–79.
[4] G. M. Weiss and F. Provost, "The effect of class distribution on classifier learning: An empirical study," Dept. Comput. Sci., Rutgers Univ., Newark, NJ, Tech. Rep. TR-43, 2001.
[5] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," Comput. Intell., vol. 20, no. 1, pp. 18–36, Feb. 2004.
[6] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in Proc. 14th Int. Conf. Mach. Learn., 1997, pp. 179–186.
[7] S. García, J. Derrac, I. Triguero, C. J. Carmona, and F. Herrera, "Evolutionary-based selection of generalized instances for imbalanced classification," Knowl.-Based Syst., vol. 25, no. 1, pp. 3–12, Feb. 2012.
[8] N. García-Pedrajas, J. A. Romero del Castillo, and D. Ortiz-Boyer, "A cooperative coevolutionary algorithm for instance selection for instance based learning," Mach. Learn., vol. 78, no. 3, pp. 381–420, Mar. 2010.
[9] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," Evol. Comput., vol. 17, no. 3, pp. 275–306, Fall 2009.
[10] L. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," Mach. Learn., vol. 51, no. 2, pp. 181–207, May 2003.
[11] N. García-Pedrajas, C. García-Osorio, and C. Fyfe, "Nonlinear boosting projections for ensemble construction," J. Mach. Learn. Res., vol. 8, pp. 1–33, May 2007.
[12] R. Barandela, J. L. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," Pattern Recognit., vol. 36, no. 3, pp. 849–851, Mar. 2003.
[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.
[14] N. García-Pedrajas, "Constructing ensembles of classifiers by means of weighted instance selection," IEEE Trans. Neural Netw., vol. 20, no. 2, pp. 258–277, Feb. 2009.
[15] A. Frank and A. Asuncion, UCI Machine Learning Repository, 2010.