A Novel Imbalance Learning Approach using in Excess and less than Sampling

L.L.SuryaPrasanthi[1]
Research Scholar, Department of Computer Science, Krishna University, Machilipatnam, India.

R.Kiran Kumar[2]
Department of Computer Science, Krishna University, Machilipatnam, India.

Kudipudi Srinivas[3]
Department of Computer Science & Engineering, V.R. Siddartha Engineering College, Vijayawada, India.

***Abstract*—In data mining, imbalance learning is a challenging task due to the intrinsic properties of the imbalance datasets. An imbalance data consists of unequal ratio instances in the classes. To address the limitations of imbalance data, we propose a novel algorithm dubbed as, In Excess Less Than (IELT) sampling technique taking into account both under sampling ad over sampling. In fact, our algorithm is capable of restructuring the original dataset at a very high conceptual level to alleviate the problems in the class imbalance. We conduct the empirical benchmark experimental setup using 15 datasets of varying class imbalance level. The proposed IELT approach performs effectively than the compared five algorithms on three evaluation metrics.**

Keywords*— *Data Mining, Knowledge Discovery, Classification, oversampling, under sampling, In Excess Less Than (IELT) sampling.

I. INTRODUCTION

Decision trees are the mathematical based algorithmic model which uses logic as the core unit for decision making. Decision tree consists of the branches and leaves. Each branch is a path of splitting the records into a narrow space and each leaf is the result of the classification of records in a specific class. There are numerous models of decision trees, which access the data and classify them in the predefined classes.

Rukshan Batuwita et al., [1] have studied the SVMs models on imbalance data learning and concluded that the learning process tends to improve the majority class and decreases the predictive ability for minority class. Rushi Longadge et al., [2] have gathered the evidence to show that a large number of existing algorithms build model to better predict majority class examples due to availability of examples and mistakenly classifies minority instances into wrong classes when imbalance dataset are applied. Kun Jiang et al., [3] have developed a hybrid algorithm GASMOTE using genetic algorithm for resample of instances in the SMOTE approach and they also used an optimal threshold for minority sampling guided by genetic algorithm.

Shaza M. Abd Elrahman et al., [4] have reviewed the latest trends in the field of class imbalance learning, which provided novel solutions to the concern issues. Bartosz Krawczyk [5] has provided a study for varied benchmark solutions for different fields in the data mining such as supervised learning, unsupervised learning, uncertainty stream learning and data with large volumes and complexity. The review of the recent works suggests that the efficiency of the decision tree reduces drastically when applied for class imbalance data sources. The reason for the reduce in performance is due to the inefficient model built with the rare instances class.

The reaming paper is distributed as given below: Section 2 presents the recent works on decision trees. Section 3 presents the main framework of the proposed IELT algorithm. Section 4 presents the details of compared algorithms and the evaluation criteria's used in the experiments. Section 5 presents the detailed experimental results and discussion. In section 6, the concluding remarks are presented with the future extension of the work.

II. CURRENT APPROACHES IN DECISION TREES

The main recent contributions in the field of decision tree are given below.

Chao Chen et al., [6] have proposed new approaches using cost sensitive learning and sampling techniques to deal the problem of class imbalance learning using random forest as the base classifier. Anne Ruiz-Gazen et al., [7] have reviewed the applicability of random forest algorithm for elevating the problem of class imbalance. They also used the logistic regression technique to forecast the resample ratio of majority and minority instances. Yuxin Peng [8] has proposed an approach for adaptively over-sampling of instances in the minority sub class and under sampling of instances in the majority sub class for better improvement of class imbalance learning measures. Gary M. Weiss et al., [9] have developed a model dubbed as Uncertainity Sampling with Biasing Consensus (USBC) which uses the ranking and weighting techniques for learning from imbalanced data sets. Yukun Chen et al., [10] have compared three methods: incorporating the misclassification costs, oversampling and under sampling. The authors determined the best technique for skewed class.

Jason Van Hulse et al., [11] have analyzed varied scenarios for sampling of data with different classifier for improvement of the skewed datasets. Jie Gu et al., [12] have proposed a hybrid model for imbalance data learning using diverse sampling techniques and random forest as the base algorithm. Vladimir Nikulin et al., [13] have proposed a model of a

classifier which performs random sampling to form balanced subsets using the available instances from both majority and minority class. The approach uses ensemble technique to adopt more than one classifier to improve the knowledge discovery process from imbalance data learning. Vladimir Nikulin et al., [14] have proposed a novel approach which selects only features with stable influence on class value to combat the problem of class imbalance.

Clearly, there are numerous different calculations which are excluded in this writing. A significant examination of the above calculations and numerous others can be accumulated from the references list

III. The Method Anticipated

This section presents the detail architecture of the proposed In Excess Less Than (IELT) approach which consists of four major modules. The detailed working principles of the IELT approach are explained below in the sub-sections.

In the initial stage of our frame work the dataset is divided into minority subset P € pi (i = 1,2,..., pnum) and majority subset N € ni(i = 1,2,..., nnum) respectively. The minority subset is the class of instances which are very less when compared to the other class in the dataset. The majority subset is the class of instances, which are more in percentage than the other class.

As the traditional algorithms efficiency drops down on imbalance data, to improve the efficiency, the dataset's majority subclass is to be under sampled or minority subclass is to be oversampled. In our proposed approach we initiated the both under sampling and oversampling strategy for the majority and minority sub classes respectively. One of the limitations of the existing oversampling algorithms is of not considering for removal of noisy and outlier instances before oversampling. Therefore, in the proposed approach before oversampling phase is started mostly misclassified instances are removed from the dataset in the form of under sampling. The technique proposed for identifying the mostly misclassified instances is by considering the nearest neighbor instances. If all the nearest neighbor instances of a particular instance are of opposite class then it implies that particular instance comes under the category of a noisy or outlier instance and can be eliminated.

The instances in the majority subset are reduced by following the below mentioned techniques; one of the technique is to eliminate the noise instances, the other technique is to find the outliers and the final technique is to find the range of weak instances for removal. The noisy and outlier instances can be easily identified by analyzing the intrinsic properties of the instances. The range of weak instances can be identified by first identifying the weak features in the majority subset. The correlation based feature selection [15] technique selects the important features by following the inter correlation between feature - feature and the inter correlation between feature and class. The features which have very less correlation are identified for elimination. The range of instances which belong to these weak features are identified for elimination from the majority subset. The number of features and instances eliminate by the correlation based feature selection technique will vary from dataset to dataset depending upon the unique properties of the dataset. The eliminated instances can boost the performance of the proposed approach in two ways:

First it will reduce the noisy and outlier instances not only from majority but also minority subset and hence improves the quality of the dataset. Second it reduces some of the outlier and noisy instances from majority subset and so reduces the imbalance nature of the dataset.

In the next phase minority subset is oversampled. The some of the synthetic instances generated are the replica of the existing instances, hybrid instances and pure artificial instances. In the final stage the fine-tuned dataset is applied to base algorithm here C4.5 [16] is considered and evaluations metric are generated.

The detailed procedure of IELT is given in the form of algorithm as follows.

Algorithm: In Excess Less Than (IELT)

Algorithm: New Predictive Model
Input: D – Data Partition, A – Attribute List
Output: A Decision Tree

Procedure:
Processing Phase:
Step 1. *Take the class imbalance data and divide it into majority and minority sub sets. Let the minority subset be P € pi* (*i* = 1,2,..., *pnum*) *and majority subset be N € ni*(*i* = 1,2,..., *nnum*).

Let us consider
m' = the number of minority nearest neighbors; *T= the whole training set*
m= the number of nearest neighbors

Step 2. *Find mostly misclassified instances pi*

pi = m'; *where m'* ($0 \leq m' \leq m$)
if $m/2 \leq m' < m$ *then pi is a mostly misclassified instance. Then remove the instances pi from the minority set.*

Let us consider
m' = the number of minority nearest neighbors

Step 3. *Find mostly misclassified instances ni*
ni = m'; *where m'* ($0 \leq m' \leq m$)
if $m/2 \leq m' < m$ *then ni is a mostly misclassified instance. Then remove the instances nifrom the majority set.*

Let us consider
m' = the number of minority nearest neighbors

Step 4. *Find noisy instances pi'*

pi' = m'; *where m'* ($0 \leq m' \leq m$)

If m'= m, i.e. all the m nearest neighbors of pi are majority examples, pi' is considered to be noise or outliers or missing values and are to be removed.

Let us consider
m' = the number of minority nearest neighbors

Step 5. *Find noisy instances ni'*

ni' = m'; where m' ($0 \le m' \le m$)
If m'= m, i.e. all the m nearest neighbors of ni are minority examples, ni' is considered to be noise or outliers or missing values and are to be removed.

Step 6. *For every pi' (i = 1,2,..., pnum') in the minority class P, we calculate its m nearest neighbors from the whole training set T. The number of majority examples among the m nearest neighbors is denoted by m'* ($0 \le m' \le m$).

If m'= m, i.e. all the m nearest neighbors of pi are majority examples, pi' is considered to be noise or outliers or missing values and are to be removed.

Step 7. *In this step, we generate s × dnum synthetic minority examples from the minority sub set, where s is an integer between 1 and k . One percentage of synthetic examples generated is replica of minority examples and other are the hybrid of minority examples.*

Selection Phase
Step 1: **begin**
Step 2: $k \leftarrow 0$,j←1.
Step 3: **Apply** CFS on subset *N*,
Step 4: Find *Fj* from N, k= number of features extracted in CFS
Step 5: **repeat**
Step 6: k=k+1
Step 7: Select the range for weak or noises instances of *Fj*.
Step 8: Remove ranges of weak attributes and form a set of major class examples N*strong*
Step 9: **Until** j = k
Step 10: Form a new dataset using *Pstrong*and *Nstrong*
Step 11:**End**

Building Predictive Model:
Step 1: *Create a node N*
Step 2: ***If*** *samples in N are of same class, C* ***then***
Step 3: *return N as a leaf node and mark class C;*
Step 4: ***If*** *A is empty* ***then***
Step 5: ***return*** *N as a leaf node and mark with majority class;*
Step 6: ***else***
Step 7: *apply C4.5*
Step 8: ***endif***
Step 9: ***endif***
Step 10: *Return N*

IV. Investigational Design and Evaluation Criteria

In order to test the strength of our method compared to existing methods, we included C4.5 [16], Reduced Error Pruning Tree (REP) [17], Classification and Regression Trees (CART) [18] Naïve Bayes Tree (NB Tree) [19] and ID3 in our experiments. Open source tool Weka is used [20] and IELT model is implemented. The 10 fold cross validation (CV) for 10 runs are used for experimental simulation. In this 10 fold CV, the dataset is divided into 10 folds and for every run, nine folds are used for training and 10^{th} fold is used for testing. The testing fold is changed to make use of all the ten folds in testing for 10 runs.

TABLE I. THE UCI DATASETS AND THEIR PROPERTIES

S.no.	Dataset	Inst	Attributes	IR
1.	Breast-cancer	286	9	2.37
2.	Breast-cancer-w	699	9	1.90
3.	Horse-colic	368	22	1.71
4.	German_credit	1,000	20	2.33
5.	Pima diabetes	768	8	1.87
6.	Heart-c	303	13	1.19
7.	Heart-h	294	13	1.77
8.	Heart-statlog	270	13	1.25
9.	Hepatitis	155	20	3.85
10.	Ionosphere	351	35	1.79
11.	Kr-vs-kp	3196	36	1.09
12.	Labor	57	17	1.85
13.	Mushroom	8124	22	1.07
14.	Sick	3772	30	15.32
15.	Sonar	208	13	1.15

The datasets used for experimental validation are obtained from UCI [21], the details are given in table 1. The accuracy is computed with the ratio of number of correctly classified to number of incorrectly classified instances. The mathematical notation for calculation of accuracy is given below in eq (i),

$$ACC = \frac{TP + TN}{TP + FN + FP + FN} \quad \text{--------- (i)}$$

Another important measure used in decision tree is the RMS Error. The Root Mean Square Error of the tree is calculated by the mean of the square error with root for the decision tree classification.

V. Experimental Results

In this section, the results of the IELT approach are compared and discussed. The results are summarized as follows.

The experimental result of the accuracy for different algorithms C4.5, REP, CART, NB Tree, ID3 on all the data sets verses proposed approach IELT are presented in table 2. From table 2 we can see that the accuracy performance of IELT model is improved on all most all the datasets. The detailed 'wins/tie/loss' are summarized in the table 5.

Table 3 shows the detailed experimental simulations of Area Under Curve (AUC) for the proposed IELT algorithm with the compared algorithms C4.5, REP, CART, NB Tree and ID3. From Table 3 we can see IELT model have performed well in terms of AUC and have achieved better performance than C4.5, ID3 and moderate improvement over REP, CART and NB Tree

TABLE II ACCURACY ON ALL THE DATASETS WITH SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE

Dataset	C4.5	REP Tree	CART	NB Tree	ID3	IELT
Breast-cancer	74.28±6.05●	69.35±5.34●	70.22±5.19●	70.99±7.94●	58.95±9.22●	92.22±3.51
Breast-cancer-w	95.01±2.73●	94.79±2.74●	94.74±2.60●	96.38±2.23●	90.62± 3.20●	98.79±1.04
Horse-colic	85.16±5.91●	84.94±5.73●	85.37±5.41●	81.71±6.39●	52.58± 8.09●	89.67±4.03
German_credit	71.25±3.17●	72.02±3.38●	73.43±4.00●	74.27±4.22●	8.94± 3.03●	84.53±2.71
Pima_diabetes	74.49±5.27●	74.46±4.39●	74.56±5.01●	75.24±5.23●	26.15± 4.31●	90.89±2.57
Heart-c	76.94±6.59○	77.02±7.24○	78.68±7.43○	80.43±6.98○	33.62±7.77●	44.39±7.79
Heart-h	80.22±7.95○	78.56±6.46○	79.02±7.18○	82.26±6.68○	27.58±7.75●	47.72±7.79
Heart-statlog	78.15±7.42○	76.15±6.71○	78.07±8.58○	79.26±8.34○	34.67±9.11●	61.20±7.61
Hepatitis	79.22±9.57●	78.62±7.07●	77.10±7.12●	80.93±9.66●	27.75±10.18●	87.50±6.43
Ionosphere	89.74±4.38●	89.46±4.56●	88.87±4.84●	89.15±5.00●	17.32± 4.79●	90.21±3.62
Kr-vs-kp	99.44±0.37●	99.01±0.55●	99.35±0.43●	97.81±2.05●	99.60±0.38●	99.70±0.33
Labor	78.6±16.58●	78.2±17.09●	80.03±16.67●	92.27±11.79●	59.33±20.60●	94.89±7.87
Mushroom	100.00±0.00	99.98±0.08●	99.95±0.09●	100.00±0.00	100.0±0.0	100.0±0.0
Sick	98.72±0.55○	98.68±0.57○	98.85±0.54○	97.82±0.76○	80.78±1.88●	83.75±1.98
Sonar	73.61±9.34●	72.69±10.19●	70.72±9.43●	77.07±9.65●	70.96±1.93●	88.02±6.13

○ Empty dot indicates the loss of IELT. ● Bold dot indicates the win of IELT;

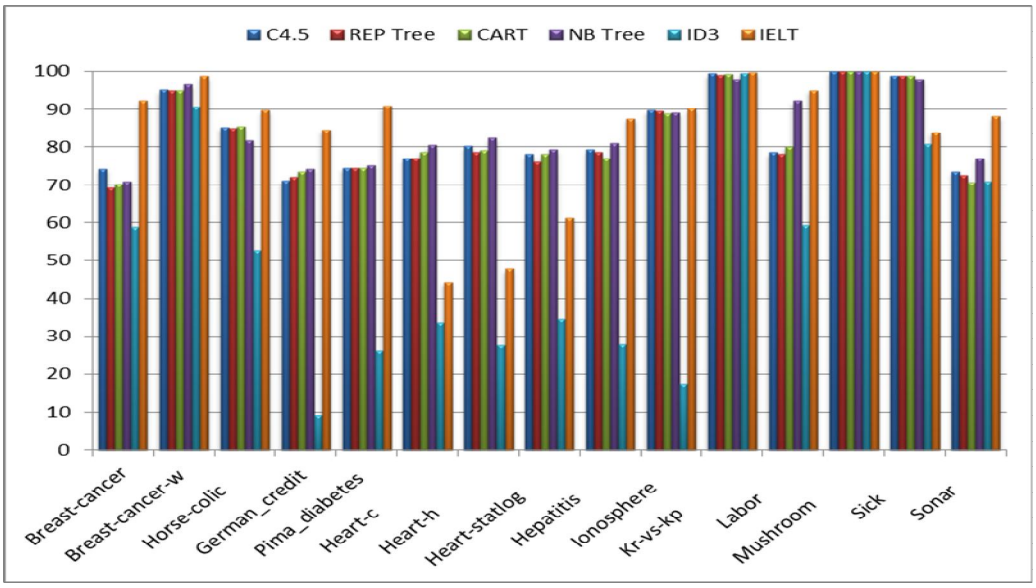

Fig. 1 Trends in accuracy for C4.5, REP, CART, NB Tree and ID3 versus IELT on UCI data sets

TABLE III. AUC ON ALL THE DATASETS WITH SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE

Dataset	C4.5	REP Tree	CART	NB Tree	ID3	IELT
Breast-cancer	0.606±0.087●	0.580±0.109●	0.587±0.110●	0.663±0.107●	0.593±0.097●	0.955±0.034
Breast_w	0.957±0.034●	0.959±0.029●	0.950±0.032●	0.986±0.015●	0.953±0.024●	0.990±0.009
Horse-colic	0.840±0.070●	0.847±0.065●	0.847±0.070●	0.859±0.070●	0.716±0.060●	0.908±0.036
German_credit	0.640±0.062●	0.712±0.053●	0.716±0.055●	0.760±0.056●	0.513±0.035●	0.887±0.020
Pima_diabetes	0.751±0.070●	0.761±0.057●	0.743±0.071●	0.804±0.055●	0.539±0.052●	0.925±0.022
Heart-c	0.769±0.082○	0.811±0.079○	0.810±0.074○	0.881±0.063○	0.573±0.088●	0.617±0.067
Heart-h	0.775±0.089○	0.826±0.074○	0.775±0.088○	0.897±0.059○	0.545±0.075●	0.607±0.057
Heart-statlog	0.786±0.094○	0.783±0.083○	0.791±0.094○	0.842±0.077○	0.591±0.084●	0.664±0.058
Hepatitis	0.668±0.184●	0.620±0.150●	0.563±0.126●	0.826±0.135●	0.474±0.043●	0.978±0.043
Ionosphere	0.891±0.060●	0.899±0.055●	0.896±0.059●	0.920±0.048●	0.738±0.064●	0.993±0.012
Kr-vs-kp	0.998±0.003○	0.998±0.002○	0.997±0.004	0.994±0.006●	0.996±0.004●	0.997±0.003
Labor	0.726±0.224●	0.768±0.233●	0.750±0.248●	0.964±0.093●	0.713±0.193●	0.977±0.057
Mushroom	1.000±0.000	1.000±0.000	0.999±0.001	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.952±0.040○	0.968±0.030○	0.954±0.043○	0.938±0.038○	0.871±0.033●	0.913±0.015
Sonar	0.753±0.113●	0.749±0.105●	0.721±0.106●	0.831±0.099●	0.498±0.013●	0.961±0.039

○ Empty dot indicates the loss of IELT. ● Bold dot indicates the win of IELT;

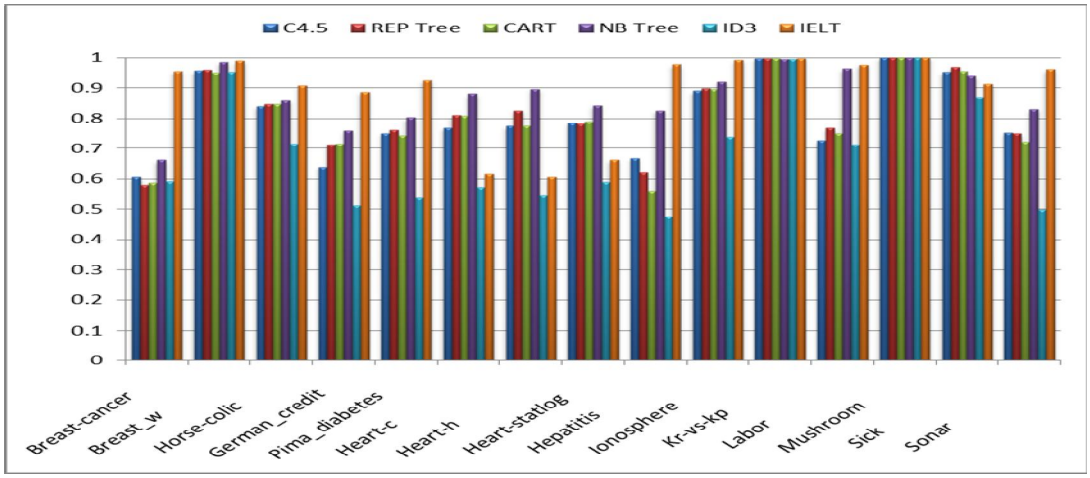

Fig. 2 Trends in AUC for C4.5, REP, CART, NB Tree and ID3 versus IELT on UCI data sets

TABLE IV. RMS ERROR ON ALL THE DATASETS WITH SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE

Dataset	C4.5	REP Tree	CART	NB Tree	ID3	IELT
Breast-cancer	0.444±0.037●	0.466±0.032●	0.458±0.039●	0.473±0.057●	0.567±0.072●	0.202±0.078
Breast_w	0.205±0.060●	0.209±0.056●	0.213±0.058●	0.169±0.062●	0.185±0.070●	0.080±0.049
Horse-colic	0.352±0.060●	0.353±0.058●	0.346±0.059●	0.379±0.072●	0.391±0.105●	0.125±0.095
German_credit	0.476±0.028●	0.441±0.025●	0.435±0.026●	0.428±0.034●	0.595±0.114●	0.078±0.057
Pima_diabetes	0.439±0.042●	0.430±0.032●	0.432±0.036●	0.417±0.037●	0.624±0.059●	0.141±0.058
Heart-c	0.281±0.039○	0.261±0.036○	0.258±0.039○	0.241±0.044○	0.398±0.058●	0.344±0.050
Heart-h	0.252±0.043○	0.251±0.033○	0.256±0.039○	0.225±0.041○	0.379±0.072●	0.274±0.059
Heart-statlog	0.429±0.077●	0.425±0.059●	0.415±0.080●	0.394±0.070○	0.598±0.101●	0.398±0.087
Hepatitis	0.404±0.096●	0.402±0.057●	0.419±0.052●	0.371±0.099●	0.510±0.221●	0.037±0.085
Ionosphere	0.299±0.081●	0.293±0.065●	0.302±0.068●	0.299±0.078●	0.050±0.131●	0.014±0.040
Kr-vs-kp	0.069±0.028●	0.090±0.026●	0.072±0.029●	0.128±0.055●	0.050±0.039●	0.042±0.036●
Labor	0.401±0.170●	0.387±0.166●	0.380±0.183●	0.200±0.163●	0.425±0.274●	0.066±0.140
Mushroom	0.000±0.000	0.005±0.014●	0.013±0.019●	0.002±0.001●	0.0±0.0	0.0±0.0
Sick	0.105±0.024●	0.106±0.023●	0.099±0.027●	0.136±0.024●	0.118±0.025●	0.087±0.032
Sonar	0.491±0.093●	0.452±0.071●	0.474±0.078●	0.434±0.098●	0.130±0.344●	0.000±0.000

○ Empty dot indicates the loss of IELT. ● Bold dot indicates the win of IELT

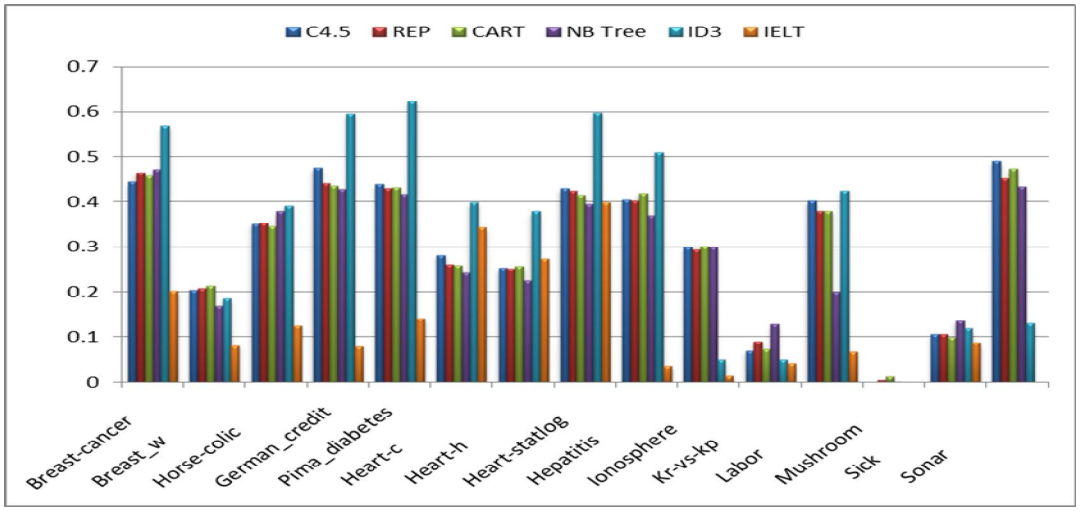

Fig. 3 Trends in RMS Error for C4.5, REP, CART, NB Tree and ID3 versus IELT on UCI data sets

The experimental result of the Root Mean Square (RMS) Error for different algorithms C4.5, REP, CART, NB Tree, ID3 on all the data sets verses proposed approach IELT are presented in table 4.From Table 4 we can see error reduction of IELT model with a substantial decrease in error on all most all the datasets. The detailed 'wins/tie/loss' are summarized in the table 5. The Figure 1-3 presents the results in the form of bar charts for easy analysis.

TABLE V. SUMMARY OF EXPERIMENTAL RESULTS FOR IELT

Results	Systems	Wins	Ties	Losses
Accuracy	IELT v/s C4.5	10	1	4
	IELT v/s REP Tree	11	0	4
	IELT v/s CART	11	0	4
	IELT v/s NB Tree	10	1	4
	IELT v/s ID3	14	1	0
AUC	IELT v/s C4.5	9	1	5
	IELT v/s REP Tree	9	1	5
	IELT v/s CART	9	2	4
	IELT v/s NB Tree	10	1	4
	IELT v/s ID3	14	1	0
RMS Error	IELT v/s C4.5	12	1	2
	IELT v/s REP Tree	13	0	2
	IELT v/s CART	13	0	2
	IELT v/s NB Tree	12	0	3
	IELT v/s ID3	14	1	0

VI. CONCLUSION

This paper presents a novel algorithm dubbed as, In Excess Less Than (IELT) sampling technique taking into account both under sampling ad over sampling. In fact, the proposed approach restructures the original imbalance dataset at a very high conceptual level to alleviate the problems in the class imbalance. We conduct the empirical benchmark experimental setup using 15 datasets of varying class imbalance level. The experimental simulations indicate that the proposed approach performs effectively than the existing approaches. In future work, we want to extend our efforts towards multi class imbalance learning.

REFERENCES

[1] RukshanBatuwita and VasilePalade,"CLASS IMBALANCE LEARNING METHODS FOR SUPPORT VECTOR MACHINES", Imbalanced Learning: Foundations, Algorithms, and Applications,. By Haibo He and Yunqian Ma, Copyright c 2012 John Wiley & Sons, Inc.

[2] RushiLongadge, Snehlata S. Dongre, Latesh Malik," Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN) Volume 2, Issue 1, February 2013 www.ijcsn.org ISSN 2277-5420.

[3] Kun Jiang, Jing Lu, Kuiliang Xia," A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE", Arab J SciEng, DOI 10.1007/s13369-016-2179-2.

[4] Shaza M. AbdElrahman and AjithAbraham,"A Review of Class Imbalance Problem"Journal of Network and Innovative Computing ISSN 2160-2174, Volume 1 (2013) pp. 332-340 © MIR Labs, www.mirlabs.net/jnic/index.html

[5] BartoszKrawczyk," Learning from imbalanced data: open challenges and future directions",ProgArtifIntell, DOI 10.1007/s13748-016-0094-0

[6] Chen, C., Com, A.L., &Liaw, A.. (2004). Using Random Forest to Learn Imbalanced Data.

[7] Anne Ruiz-Gazen, and Nathalie Villa, "Storms prediction : Logistic regression vs random forest for unbalanced data", Cases Studies in Business, Industry and Government Statistics (CSBIGS), vol. 1, n. 2, 2007, pp. 91–101.

[8] YuxinPeng," Adaptive Sampling with Optimal Cost for Class-Imbalance Learning", AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Pages 2921-2927.

[9] Gary M. Weiss, Kate McCarthy, and BibiZabar," Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?", *Proceedings of the 2007 International Conference on Data Mining*, CSREA Press, 35-41.

[10] Yukun Chen, Subramani Mani "Active Learning for Unbalanced Data in the Challenge with Multiple Models and Biasing", JMLR:Workshop and Conference Proceedings 16 (2011) 113-126Workshop on Active Learning and Experimental Design.

[11] Jason Van Hulse, Taghi M. Khoshgoftaar, Amri Napolitano "Experimental Perspectives on Learning from Imbalanced Data", Proceedings of the 24 th International Conference on Machine Learning, Corvallis, OR, 2007

[12] JieGu, Yuanbing Zhou, and XianqiangZuo "Making Class Bias Useful: A Strategy of Learning from Imbalanced Data", in Proceedings of Intelligent Data Engineering and Automated Learning - IDEAL 2007, 8th International Conference, Birmingham, UK, December 16-19, 2007

[13] Vladimir Nikulin, Geoffrey J. McLachlan "Classification of Imbalanced Marketing Data with Balanced Random Sets", JMLR: Workshop and Conference Proceedings 7: 89-100.

[14] Vladimir Nikulin, Geoffrey J. McLachlan, and Shu Kay Ng "Ensemble Approach for the Classification of Imbalanced Data", A. Nicholson and X. Li (Eds.): AI 2009, LNAI 5866, pp. 291–300, 2009.

[15] Hall MA (1998) Correlation-based feature subset selection for machine learning. PhD Thesis.

[16] J. R Quinlan,.(1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos.

[17] J. Quinlan. "Induction of decision trees, Machine Learning", vol. 1, pp. 81C106, 1986.

[18] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

[19] Ron Kohavi: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: Second International Conference on Knoledge Discovery and Data Mining, 202-207, 1996.

[20] Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

[21] HamiltonA. Asuncion D. Newman. (2007). *UCI Repository of Machine Learning Database* (School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html