# A New Hybrid Sampling Approach for Classification of Imbalanced Datasets

Anantaporn Hanskunatai
Advanced Artificial Intelligence Research Laboratory
Department of Computer Science
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang, Bangkok, 10520, Thailand
e-mail: anantaporn.ha@kmitl.ac.th

***Abstract*—Nowadays it is an era of data driven. Many organizations around the world including bank, industry, commercial, and medical intend to extract knowledge from a huge of data. But in the real-word datasets, most of them occur class imbalance problems. This paper presents a new algorithm to handle an imbalanced classification. The proposed technique is a hybrid sampling approach which is the combination of a well know oversampling algorithm called SMOTE and the undersampling technique by removing the ambiguous instances from the majority class instances. The experimental results show that the new hybrid sampling method yields the better predictive performance in term of *F-measure* when compare with other sampling techniques. In addition, it can improve f-measure up to 59.73% and 412.26% when compare with the original dataset based on decision tree learning and naïve bayes classifiers respectively.**

***Keywords-imbalanced dataset; SMOTE; DBSCAN; hybrid sampling; decision tree; naïve bayes***

## I. INTRODUCTION

In recently, there are many organizations such as commercial, medical, finance, marketing, and health care use data mining to extract important and interesting knowledge or patterns from their amount of data. However, the predictive performance of the model is depended on the training dataset. The more volume of training data the more accurate in classification performance of the model. In real world problems of binary classification, there are imbalanced between the number of positive and negative instances in the training set. For an example of diagnosis application, the number of people who infect or be a cancer (positive instance) is smaller than the number of normal people. In theoretically of machine learning, some learning algorithms such as decision tree or naïve bayes are sensitive to training set. This means that if the training dataset is slightly change the generated model will be different. In addition, these learning algorithms are probabilistic models. Thus, their likely to predict the results as the majority class (negative class).

In few years ago, there are many researchers have been proposed various techniques for solving the class imbalanced problem. At the summary, there are three main approaches for handle the imbalanced classification problem which are data sampling, feature selection, and ensemble. The first solution is applying a sampling technique for balancing the number of instances between positive and negative class. In [1] the new undersampling algorithm called DE was present. The concept of DE is using evolutionary computation for finding a subset of majority class that is imperfect for classification. In addition, the oversampling method based on immune network theory was proposed in [2]. The second approach is a feature selection technique. In [3], decision tree was applied as filter-based for feature selection in the data preprocessing step for class imbalance from Santander Bank. The last approach for enhancing the classification performance in class imbalance problem is an ensemble technique. This method was used in many researches based on incrementally learning technique such as Learn++ [4-7]. Furthermore, the combination between feature selection technique and ensemble was proposed in [8].

## II. THE PROPOSED HYBRID SAMPLING METHOD

The concept of the proposed hybrid sampling method is to eliminate unimportant instances of negative class (or majority class) and simulate positive instances (or minority class) in the overlap area between positive and negative instances. Fig. 1 shows the framework of the proposed sampling method.

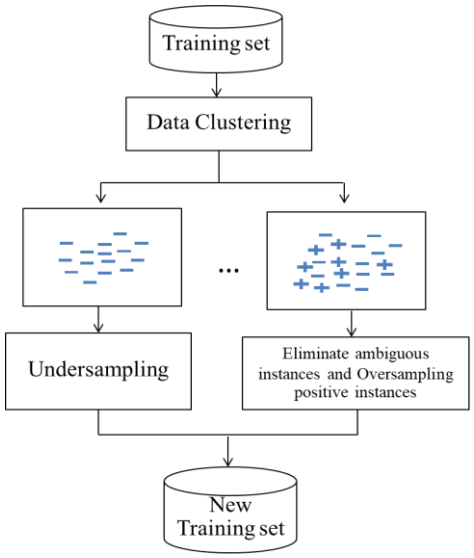


Figure 1. Framework of the proposed hybrid sampling technique.

978-1-5386-6350-9/18/$31.00 ©2018 IEEE

There are three main parts of the proposed algorithm; grouping, sampling, and gathering. In the grouping process, DBSCAN [9] was applied to find the clusters (line 1 in Fig. 2). Then each cluster is considered for sampling individually. If all members in the cluster are negative class, 50% of all instances nearest to the center of the cluster will be removed (line 5-8 in Fig. 2). If members in the cluster are positive and negative instances, the algorithm will find overlap area called set S (line 10-12 in Fig.2). The overlap area consists of instances that misclassify by *k*NN with *k*=3. Considering the overlap area, the ambiguous instance is defined as the instance misclassify by *k*NN with *k*=5. Then all ambiguous instances are eliminated from set S. After that, only remaining positive instances in set S are sent to SMOTE algorithm [10] for oversampling of a positive class. Finally, the new training set is formed by integrating original positive instances including synthetic instances from SMOTE and remaining negative instances (line 23 in Fig. 2).

**Algorithm:** Hybrid Sampling
**Input:** D: All training set with positive instances and negative instances
**Output:** P: New training set from hybrid sampling.

1. [Cluster] = DBSCAN(D) // Cluster is a set of clusters generated by the DBSCAN algorithm
2. $T_{pos}$={ }; $T_{neg}$={ };
3. **For** i = 1 to n // n is the number of clusters generated by DBSM
4. T={ }; $T_{SMOTE}$= { }; S={ };
5. **If** all members in $Cluster_i$ are negative instance **then**
6. Remove 50% of instances from the $Cluster_i$
7. T = remaining instances in the $Cluster_i$
8. $T_{neg} = T_{neg} \cup T$
9. **Else if** members in $Cluster_i$ are positive and negative instances **then**
10. **For** j = 1 to m1 // m1 is a total instances in the $Cluster_i$
11. **If** $instance_j$ is misclassified by kNN **then**
12. $S = S \cup instance_j$
13. **End for**
14. Remove ambiguous instances from set S
15. T=all remaining positive instances in set S
16. $T_{SMOTE}$=positives instances generated by SMOTE(T)
17. T = { };
18. T= $T_{SMOTE} \cup$ positive instances in the $Cluster_i$ outside set S
19. $T_{neg} = T_{neg} \cup$ all negative instances in the $Cluster_i$ including negative samples in set S
20. $T_{pos} = T_{pos} \cup T$
21. **End if**
22. **End for**
23. Return $P = T_{pos} \cup T_{neg}$

Figure 2. Pseudo code of the proposed hybrid sampling method.

## III. EXPERIMENTAL RESULTS

The scope of this work is limited to a binary classification problem. Twelve datasets downloaded from KEEL [11] were used in the experiment for evaluating the performance of the models after applied various sampling methods including the proposed hybrid sampling algorithm. The characteristic of the datasets is shown in Table I.

TABLE I. DATASETS CHARACTERISTIC

| Dataset Name | #Attributes | #Examples | IR |
|---|---|---|---|
| glass1 | 9 | 214 | 1.82 |
| wiscosin | 9 | 683 | 1.86 |
| glass0 | 9 | 214 | 2.06 |
| yeast1 | 8 | 1484 | 2.46 |
| haberman | 3 | 306 | 2.78 |
| vehicle2 | 18 | 846 | 2.88 |
| vehicle1 | 18 | 846 | 2.9 |
| new-thyroid1 | 5 | 215 | 5.14 |
| new-thyroid2 | 5 | 215 | 5.14 |
| ecoli2 | 7 | 336 | 5.46 |
| glass6 | 9 | 214 | 6.38 |
| yeast3 | 8 | 1484 | 8.1 |

The Imbalance ratio (IR) (the last column in the Table I) is a number of positive instances compare with negative instances which is calculated by (1) where n is the number of negative instances and m is the number of positive instances.

$$\text{imbalance ratio} = \frac{n}{m} \quad (1)$$

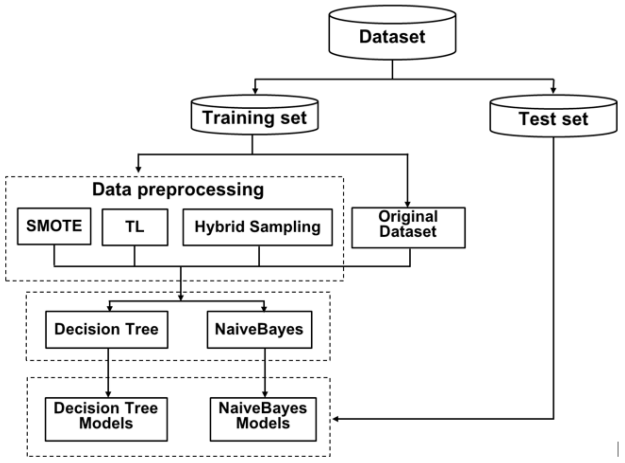


Figure 3. Design of the experiment.

For the experimental design, the datasets were separated into two parts; training and test sets by using *5*-fold cross validation technique. At the first step, the training set was dealt with data preparation step. In this process, various algorithms of data sampling such as SMOTE, Tomek Links

(TL) [12], and the proposed hybrid sampling method were used to sampling instances in the training set for balancing ratio between the positive and negative instances. At the summary, there were four training datasets generated from three different sampling methods including an original training data. To evaluate the performance of the proposed hybrid sampling method with the other sampling techniques, two popular machine learning algorithms; decision tree and naïve bayes were used to construct the classification models and evaluated the classification performance with the same test set. The framework of the experimental design is illustrated in Fig 3.

In general, accuracy is a basic metric to evaluate the predictive performance of a classifier. But when a dataset is imbalance, *F-measure* is an important measurement for evaluating the performance of the model. The formula of *F-measure* is calculated by (2). *Recall* refers to how well a classifier can regcognize a positive class while *precision* evaluates how accurate of a classifier in corrrectly classify positive instances. *Precision* is a proportion of the positive instances that correctly classified as positive instances (*TP*) and instances classified by a model as positive class (*TP*+*FP*).

$$F\text{-}measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (2)$$

Where *recall* and *precision* are calculated by (3) and (4)

$$recall = \frac{TP}{TP+FN} \quad (3)$$

$$precision = \frac{TP}{TP+FP} \quad (4)$$

*TP* is the number of positive instances correctly predicted, *FP* is the number of negative instances incorrectly predicted, and *FN* is the number of positive instances incorrectly predicted.

Table II and Table III show the experimental results in accuracy of two machine learning algorithms decision tree and naïve bayes. Each table shows the comparison of different sampling techniques used in the data preprocessing step. As shown in Table II and Table III, the accuracies of the proposed hybrid sampling technique are the best in average accuracy when compared with other sampling methods including the original dataset in both decision tree and naïve bayes classifiers. In addition, the experimental results indicate that the decision tree model has a higher average accuracy than the naive bayes model in the original dataset and all sampling techniques. Considering the results in Table II and Table III again, the best algorithm that yields the highest average accuracy is the proposed hybrid sampling. The second best of average accuracy for decision tree model is SMOTE as the same time it provides the wost predictive in the naïve bayes model. For TL, it is yields the third in average accuracy in the naïve bayes model but gives the lowest average accuracy in the decision tree model.

TABLE II. ACCURACY OF DECISION TREE MODELS

| Dataset | Accuracy of Decision Tree Models (%) | | | |
|---|---|---|---|---|
| | Original | SMOTE | TL | Hybrid Sampling |
| glass1 | 70.13 | **77.11** | 74.33 | 73.83 |
| wiscosin | 94.73 | 94.88 | 96.05 | **96.05** |
| glass0 | 81.75 | 80.35 | 78.53 | **85.50** |
| yeast1 | 73.65 | 71.29 | 70.01 | **76.08** |
| haberman | 66.01 | 66.32 | 65.00 | **72.85** |
| vehicle2 | 95.27 | **95.63** | 95.15 | 95.51 |
| vehicle1 | 74.11 | 74.58 | 72.45 | **78.96** |
| new-thyroid1 | 94.42 | **97.21** | 94.42 | 96.74 |
| new-thyroid2 | 96.28 | 95.81 | 93.49 | **97.21** |
| ecoli2 | 91.67 | 91.37 | 92.86 | **93.76** |
| glass6 | 93.01 | 92.08 | 93.47 | **97.20** |
| yeast3 | 93.67 | 94.27 | 93.20 | **94.88** |
| Average | 85.39 | 85.91 | 84.91 | **88.21** |

TABLE III. ACCURACY OF NAÏVE BAYES MODELS

| Dataset | Accuracy of Naïve Bayes Models (%) | | | |
|---|---|---|---|---|
| | Original | SMOTE | TL | Hybrid Sampling |
| glass1 | 60.27 | 58.88 | 59.35 | **70.55** |
| wiscosin | 96.49 | 96.49 | 95.75 | **97.37** |
| glass0 | 63.55 | 63.55 | 63.55 | **81.81** |
| yeast1 | 70.28 | 63.00 | 67.11 | **70.42** |
| haberman | 75.16 | **76.12** | 73.83 | 69.59 |
| vehicle2 | 80.38 | 76.59 | 79.91 | **88.65** |
| vehicle1 | 68.44 | 67.49 | 67.26 | **69.62** |
| new-thyroid1 | 97.21 | 96.74 | 96.74 | **98.14** |
| new-thyroid2 | 97.67 | 96.74 | 97.21 | **98.60** |
| ecoli2 | 93.46 | 93.17 | 94.36 | **95.84** |
| glass6 | 93.90 | 93.90 | 92.95 | **95.30** |
| yeast3 | 89.69 | 89.02 | 90.84 | **95.08** |
| Average | 82.21 | 80.97 | 81.57 | **85.91** |

Table IV and Table V show the experimental results of *F-measure* for the decision tree and naive bayes models. The proposed hybrid sampling algorithm performs the best in average *F-measure* not only in the decision tree model but also yields the best in the naïve bayes model. TL and SMOTE are the second best in the decision tree and naïve bayes models respectively while the original dataset is the worst for both two machine learning models (decision tree and naïve bayes). These results imply that all sampling techniques can improve the performance of positive (minority) class prediction, especially for the proposed hybrid sampling technique.

TABLE IV. F-MEASURE OF DECISION TREE MODELS

| Dataset | F-measure of Decision Tree Models | | | |
|---|---|---|---|---|
| | Original | SMOTE | TL | Hybrid Sampling |
| glass1 | 0.540 | 0.659 | **0.683** | 0.636 |
| wiscosin | 0.925 | 0.666 | **0.945** | 0.943 |
| glass0 | 0.704 | 0.609 | 0.696 | **0.770** |
| yeast1 | 0.528 | **0.734** | 0.551 | 0.546 |
| haberman | 0.287 | 0.397 | 0.439 | **0.458** |
| vehicle2 | 0.909 | 0.901 | 0.909 | **0.915** |
| vehicle1 | 0.482 | **0.897** | 0.519 | 0.574 |
| new-thyroid1 | 0.836 | 0.55 | 0.833 | **0.892** |
| new-thyroid2 | 0.886 | 0.909 | 0.819 | **0.912** |
| ecoli2 | 0.700 | **0.918** | 0.771 | 0.797 |
| glass6 | 0.734 | 0.524 | 0.746 | **0.890** |
| yeast3 | 0.703 | 0.689 | 0.720 | **0.784** |
| Average | 0.686 | 0.704 | 0.719 | **0.760** |

TABLE V. F-MEASURE OF NAIVE BAYES MODELS

| Dataset | F-measure of Naïve Bayes Models | | | |
|---|---|---|---|---|
| | Original | SMOTE | TL | Hybrid Sampling |
| glass1 | 0.645 | 0.632 | 0.601 | **0.655** |
| wiscosin | 0.954 | 0.957 | 0.952 | **0.962** |
| glass0 | 0.644 | 0.637 | 0.611 | **0.765** |
| yeast1 | 0.169 | 0.391 | 0.357 | **0.473** |
| haberman | 0.193 | 0.335 | 0.293 | **0.452** |
| vehicle2 | 0.760 | **0.782** | 0.752 | 0.778 |
| vehicle1 | 0.505 | **0.517** | 0.510 | 0.511 |
| new-thyroid1 | 0.973 | 0.950 | **0.973** | 0.947 |
| new-thyroid2 | **1.000** | 0.975 | **1.000** | 0.962 |
| ecoli2 | 0.747 | 0.777 | 0.797 | **0.867** |
| glass6 | 0.758 | 0.732 | 0.776 | **0.822** |
| yeast3 | 0.150 | 0.658 | 0.295 | **0.768** |
| Average | 0.625 | 0.695 | 0.660 | **0.747** |

Table VI shows the improvement of the accuracy based on the proposed hybrid sampling technique for the decision tree and naïve byes models. In case of the decision tree models, the proposed algorithm can improve the accuracy up to 10.36% in original dataset, 9.85% in SMOTE, and 12.08% in TL on the Haberman dataset. For naïve bayes models, the hybrid sampling technique can enhance the accuracy up to 28.73% for all technique including the original dataset on the Glass0 dataset.

Table VII summarizes the improvement of the *F-measure* based on the proposed hybrid sampling method. Considering to the decision tree models, the proposed hybrid sampling technique can enhance the *F-measure* in the original dataset, SMOTE, and TL up to 59.73% on the Haberman dataset, 69.91% on the Glass6 dataset, and 19.34% on the Glass6 dataset respectively. For the naïve bayes models, the proposed algorithm dramatically improves the prediction performance in terms of *F-measure* up to 412.26% in the original dataset on the Yeast3 dataset, 160.47% in TL on the Yeast3 dataset, and 34.88% in SMOTE on the Haberman dataset.

TABLE VI. ACCURACY IMPROVEMENT WITH THE NEW HYBRID SAMPLING METHOD

| Dataset | Decision Tree Models | | | Navie Bayes Models | | |
|---|---|---|---|---|---|---|
| | Original | SMOTE | TL | Original | SMOTE | TL |
| glass1 | 5.28 | -4.25 | -0.67 | 17.06 | 19.83 | 18.88 |
| Wiscosin | 1.39 | 1.23 | 0.00 | 0.91 | 0.91 | 1.69 |
| glass0 | 4.59 | 6.41 | 8.88 | 28.73 | 28.73 | 28.73 |
| yeast1 | 3.29 | 6.71 | 8.67 | 0.19 | 11.77 | 4.93 |
| Haberman | 10.36 | 9.85 | 12.08 | -7.41 | -8.58 | -5.75 |
| vehicle2 | 0.25 | -0.13 | 0.38 | 10.29 | 15.75 | 10.94 |
| vehicle1 | 6.54 | 5.87 | 8.98 | 1.72 | 3.15 | 3.51 |
| new-thyroid1 | 2.46 | -0.48 | 2.46 | 0.96 | 1.45 | 1.45 |
| new-thyroid2 | 0.97 | 1.46 | 3.98 | 0.96 | 1.93 | 1.43 |
| ecoli2 | 2.28 | 2.62 | 0.97 | 2.54 | 2.86 | 1.57 |
| glass6 | 4.50 | 5.56 | 3.99 | 1.50 | 1.50 | 2.53 |
| yeast3 | 1.29 | 0.65 | 1.80 | 6.01 | 6.81 | 4.67 |
| Average | 3.60 | 2.96 | 4.29 | 5.29 | 7.17 | 6.21 |

TABLE VII. F-MEASURE IMPROVEMENT WITH THE NEW HYBRID SAMPLING METHOD

| Dataset | Decision Tree Models | | | Navie Bayes Models | | |
|---|---|---|---|---|---|---|
| | Original | SMOTE | TL | Original | SMOTE | TL |
| glass1 | 17.87 | -3.42 | -6.81 | 1.51 | 3.59 | 8.94 |
| wiscosin | 1.93 | 41.57 | -0.23 | 0.89 | 0.57 | 1.10 |
| glass0 | 9.43 | 26.50 | 10.69 | 18.83 | 20.14 | 25.25 |
| yeast1 | 3.36 | -25.65 | -0.96 | 179.67 | 20.88 | 32.39 |
| haberman | 59.73 | 15.47 | 4.43 | 134.12 | 34.88 | 54.21 |
| vehicle2 | 0.64 | 1.54 | 0.64 | 2.40 | -0.48 | 3.49 |
| vehicle1 | 19.00 | -36.05 | 10.52 | 1.21 | -1.14 | 0.22 |
| new-thyroid1 | 6.69 | 62.17 | 7.08 | -2.71 | -0.35 | -2.71 |
| new-thyroid2 | 2.94 | 0.34 | 11.37 | -3.83 | -1.37 | -3.83 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ecoli2 | 13.88 | -13.17 | 3.39 | 16.08 | 11.60 | 8.80 |
| glass6 | 21.29 | 69.91 | 19.34 | 8.41 | 12.26 | 5.89 |
| yeast3 | 11.58 | 13.84 | 8.94 | 412.26 | 16.78 | 160.47 |
| Average | 14.03 | 12.75 | 5.70 | 64.07 | 9.78 | 24.52 |

## IV. CONCLUSION

This paper presents a new hybrid sampling method for solving the class imbalanced problem. The concept of the proposed algorithm is applying SMOTE for oversampling the minority class instances and undersampling the majority class instances by removing ambiguous instances. The predictive performance of the proposed method was evaluated in terms of accuracy and *F-measure* metrics, and compared with the other two sampling methods: SMOTE and TL, including the original dataset (no data sampling). The experimental results show that the proposed technique yields the highest predictive performance both in accuracy and *F-measure*.

## REFERENCES


[1] C. Qiu, L. Jiang, and G. Kong, "A Differential Evolution-Based Method for Class-Imbalanced Cost-Sensitive Learning," International Joint Conference on Neural Networks, IJCNN 2015, October, 2015, doi: 10.1109/IJCNN.2015.7280419.

[2] X. Ai, J. Wu, Z.Cui and et.al, "Enrich the Data Density of Cluster for Imbalanced Learning Using Immune Representatives," Proc. 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), January 2017, doi: 10.1109/ICSEE. 2016.7806162.

[3] H. Liu and M. Zhou, "Decision Tree Rule-based Feature Selection for Large-scale Imbalanced Data," Proc. Wireless and Optical Communication Conference (WOCC), May 2017, doi: 10.1109/ WOCC.2017.7928973.

[4] R. Razavi-Far and M. Saif, "Ensemble of Extreme Learning Machines for Diagnosing Bearing Defects in Non-stationary Environments Under Class Imbalance Condition," Proc. 2016 IEEE Symposium Series on Computational Intelligence (SSCI), February 2017, doi: 10.1109/SSCI.2016.7849967.

[5] G. Ditzler and R. Polikar, "An Ensemble based Incremental Learning Framework for Concept Drift and Class Imbalance, " Proc. The 2010 International Joint Conference on Neural Networks (IJCNN), October 2010, doi: 10.1109/IJCNN.2010.5596764.

[6] G. Ditzler and R. Polikar, "Incremental learning of Concept Drift from Streaming Imbalanced Data,", IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 10, pp. 2283-2301, 2013, doi: 10.1109/TKDE.2012.136.

[7] G. Ditzler, R. Polikar, and N. Chawla, "An Incremental Learning Algorithm for Non-Stationary Environments and Class Imbalance," Proc. International Conference on Pattern Recognition (ICPR), October 2010, pp. 2997-3000, doi: 10.1109/ICPR.2010.734.

[8] H. Yin, K. Gai, and Z. Wang, "A Classification Algorithm Based on Ensemble Feature Selections for Imbalanced-Class Dataset," Proc. 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), July 2016, doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.76.

[9] M. Ester, H. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", Second International Conference on Knowledge Discovery and Data Mining, pp. 226-231, 1996.

[10] N. V. Chawla, L. O. Hall, K.W. Bowyer, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[11] J. Alcal´a-Fdez, A. Fern´andez, J. Luengo, J. Derrac, S. Garc´ıa, L. S´anchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," J. Mult. Valued Logic Soft Comput., vol. 17, no. 2–3, pp. 255–287, 2011.

[12] Tomek, I., "Two modifications of CNN," IEEE Trans. Systems, Man and Cybernetics, vol. SMC-6, pp. 769-772, Nov. 1976.