A Novel Heuristic Method for Misclassification Cost Tuning in Imbalanced Data

Anusara Hirunyawanakul, Nittaya Kerdprasop, and Kittisak Kerdprasop

***Abstract*—Currently, one of the most challenging problem in machine learning and data mining is the data imbalance problem. Many techniques and methods are researched and proposed to solve this problem. Fundamental solution is data balancing with under-sampling and over-sampling techniques. However, these conventional methods might be suffered from the potential loss of useful information leading to the generation of useless patterns. Therefore, the techniques that avoid adjusting the sample size of data are more interesting. One of such technique is misclassification cost adjustment. This paper focuses on improving the performance of classification model built from the misclassification cost adjustment technique by proposing the novel heuristic method. Our proposed method uses a heuristic based on the experience of practitioner working on many manufacturing data. The heuristic employs the relation between misclassification cost, imbalance ratio and a constant factor "e" (Euler's number). The experiment has been operated on 56 real-world datasets with various number of attributes and different degrees of imbalance ratio. The results confirm that our novel heuristic method can help improving the performance of the classification model. On datasets with high imbalance ratio, our method shows the improvement rate of AUC up to 29%.**

***Index Terms*—Misclassification cost, imbalance data, classification, decision tree learning.**

I. Introduction

Data mining and machine learning are very popular and extensively used in several areas. The problem that has been reported as one of the most often found in this field is the imbalance ratio problem. Class imbalance data problem has been reported to occur in a wide variety of real world domains, such as facial age approximation [1], detecting oil spills from satellite images [2], anomaly detection [3], fraudulent credit card transactions detection [4], software error prediction [5], and pattern recognition on image annotation [6].

Most traditional algorithms, such as decision trees [7]–[9], k-nearest neighbors [10], [11], focus on generating the models that provide the highest overall accuracy and the minority data is always ignored [12]–[14]. However, in some cases the minority class instances may have so high significance and importance that they should not be ignored by the classification algorithms. Thus, data-preprocessing steps for balancing instances between classes are needed. One of the most popular methods for class rebalancing is data sampling [15]–[18]. However, under-sampling may eliminate the important data of the majority class. While over-sampling methods may alter the original class distribution. Moreover, increasing the minority class instances may generate the useless data and misleading the classification result. The cost-sensitive learning or misclassification cost adjustment seems to be the efficient way to solve the class imbalance problems [19]–[21].

The technique that we discovered in one field may show the good result in other fields and this paper is one of them. The technique that we introduce in this paper is extracted from the experience of researchers while had been working in the manufacturing companies and already proved with the datasets which are collected from production line database of Computer's component manufacturing. This method can help the expertise engineers to achieve the optimal of "true positive rate" in a shorter time.

This paper is used that novel method to apply on worldwide 56 datasets with the various fields like citizen data, wine quality data, card game data, medical/scientific experimental data etc. With these datasets, we separate them into 2 groups: the low imbalance ratio group and the high imbalance ratio group. The model performance can be significantly improved by our novel heuristic method especially in case of high imbalance ratio group. The remaining of this paper is organized as follows. Section II is Theory and Literature review. Section III is the Material and Method explaining the novel heuristic method and how to calculate the heuristic value. Section IV is the research workflow and research framework. Section V presents the experimental results. Section VI is the conclusion of this paper and the recommendation is presented in Section VII.

II. Background Theory and Literature Review

A. Decision Tree

Decision tree is a well-known and one of the most employed technique to generate classifier [22]. Decision tree has 3 important parts: a root node, leaf nodes, and branches to connect nodes. The root node is the origin node of the tree, and both root and other internal nodes consist of condition or criteria to be considered before selecting a branch to traverse. Each branch is a connection line between nodes. Leaf node is a final solution for a specific classification problem.

The tree building process starts with all the training data in the root node. A first split is made using a predictor variable to segment data into 2 or more child nodes, depending on the possible values of the predictor variable. The terminal node is the node that cannot be further split, and the predictions are made from the terminal nodes. To use a decision tree to make

Manuscript received August 20, 2018; revised October 12, 2018.

The authors are with School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: Anusara.hi@gmail.com, nittaya@sut.ac.th, kerdpras@sut.ac.th).

doi: 10.18178/ijmlc.2018.8.6.746

a prediction, the split decisions are followed until a terminal node is reached.

Decision trees are always mentioned as popular tools for presenting a decision-making process [23], because they are easy for understanding with the clearly graphic. But building efficient decision trees from data is quite complicated. The classical method such as ID3, developed by Quinlan [24]–[26], takes a table of examples as input, where each example consists of a collection of attributes, together with a class. And then, induces a decision tree, where each node is a test on an attribute, each branch is the outcome of that test. The last branching step leads to one of the leaf nodes consisting of the class value to which the example, when following that path, belongs. With the continuous development and improvement, many algorithms such as C4.5 and C5.0 [27] are developed to focus on how to build a decision tree efficiently based on several criteria of consideration [28].

In this research, we use the C5.0 as an algorithm to build model because it has been shown the very satisfying performance compared to other algorithms. Besides the easy-to-understand which is the strongest point of the decision tree, the robustness is also another advantage that makes decision tree popular. It has the ability to be applied with many types of data, fast in prediction, and no need for the assumption on variable distribution [29].

B. Imbalanced Data

Data imbalance is often reported as a problem to reduce classification efficiency in traditional learning algorithms. In classification task, imbalanced data problem occurs when the samples size from the majority class is heavily higher than minority class, and the minority class is usually misclassified by such classification models [30], [31]. Thus, methods to balance the skewed data, such as under-sampling and over-sampling, have been used to tackle the problem. However, under-sampling may drop some potentially useful information, while over-sampling may be the cause of another problem like overfitting [32], [33]. Therefore, it is reasonable to develop the algorithm without conversion from imbalanced data into balanced ones by introducing extra information or removing the original information. The misclassification cost adjustment or cost-sensitive learning is the answer.

The cost-sensitive learning algorithm is developed based on the assumption that the positive minority class is expected to be more important than the majority negative class. Thus, instances in positive class have been weighted with more value than those in negative class. The weighting scheme is based on the misclassification cost adjustment occurred during the iterative model assessment process. The difficulty of this method is finding a proper value for misclassification cost that should be adjusted. The optimal goal is adjusting with the value that results in the highest classification performance on both classifying the minority and majority classes. Unfortunately, a suitable value of misclassification cost comes from many times of trial and run the model repeatedly to see the satisfied result.

C. Confusion Matrix

Confusion matrix [34] is a table that is normally used as a tool for computing performance of a classification model. The key function of this table is to present a comparison between "Predicted Labels" from model and "Actual Labels" from the ground truth. Fig. 1 shows the example of classification outcome of data instances from two groups: "Positive" and "Negative".

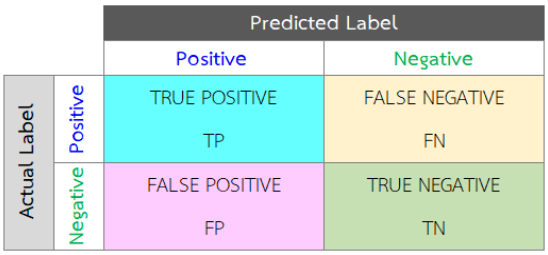

Fig. 1. Example of confusion matrix.

- *True Positive (TP):* The number of instances that a model predicts correctly such that the "Actual Labels" is Positive and "Predicted Labels" is Positive as well.
- *True Negative (TN):* The number of instances that a model predicts correctly such that the "Actual Labels" is Negative and "Predicted Labels" is Negative as well.
- *False Positive (FP):* The number of instances that a model predicts incorrectly such that the "Actual Labels" is Negative but "Predicted Labels" is Positive.
- *False Negative (FN):* The number of instances that a model predicts incorrectly such that the "Actual Labels" is Positive but "Predicted Labels" is Negative.

True Positive Rate (TPR), or Sensitivity, measures the proportion of actual positive data instances that are correctly identified. The calculation of TPR is shown in equations 1 and 2.

$$TPR = TP / (TP + FN) \quad (1)$$

or

$$TPR = TP / (\text{All actual positive instances}) \quad (2)$$

False Positive Rate (FPR) is a metric for measuring the error of classification. It is calculate with the equations 3 and 4.

$$FPR = FP / (FP + TN) \quad (3)$$

or

$$FPR = FP / (\text{All actual negative instances}) \quad (4)$$

D. Performance Evaluation

In classification, there are various measurement methods for evaluating the performance of classification models. Receiver Operating Characteristic curve (ROC) is the visualization to represent the relation of the false positive rate (FPR) against the true positive rate (TPR) by plotting graphs with TPR on the Y-axis and FPR on the X-axis. The performance of a classifier is presented by ROC curve. If it lies in the upper left of the square that means good performance.

AUC or area under the ROC curve [35], [36] is the popular measure for evaluating the performance of a classification model with binary classes. AUC provides a value description for the performance of the ROC curve. AUC is a portion of

the area inside the square of unit (Fig. 2). So, its value must be in the range of 0 and 1, and usually higher than 0.5.

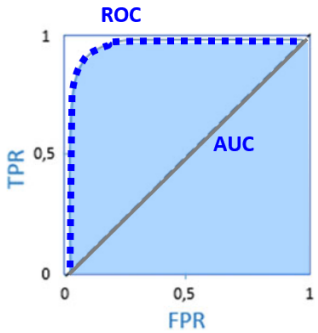

Fig. 2. Example of ROC Curve.

III. Materials and Method

A. Datasets of Research

TABLE I: The 34 Datasets of "Low Imbalance Ratio" Showing Numbers of Majority Class and Minority Class

Group Low Imbalance Ratio						
No.	Dataset	IR	# Attr.	# Ins.	# Major	# Minor
1	glass-0-1-6_vs_5	19.4	9	184	175	9
2	abalone9-18	16.4	8	731	689	42
3	page-blocks-1-3_vs_4	15.9	10	472	444	28
4	ecoli4	15.8	7	336	316	20
5	glass4	15.5	9	214	201	13
6	yeast-1_vs_7	14.3	7	459	429	30
7	shuttle-c0-vs-c4	13.9	9	1,829	1,706	123
8	ecoli-0-1-4-6_vs_5	13.0	6	280	260	20
9	cleveland-0_vs_4	12.6	13	177	164	13
10	ecoli-0-1-4-7_vs_5-6	12.3	6	332	307	25
11	glass2	11.6	9	214	197	17
12	glass-0-1-4-6_vs_2	11.1	9	205	188	17
13	ecoli-0-1_vs_5	11.0	6	240	220	20
14	glass-0-6_vs_5	11.0	9	108	99	9
15	led7digit-0-2-4-5-6-7-8-9_vs_1	11.0	7	443	406	37
16	ecoli-0-1-4-7_vs_2-3-5-6	10.6	7	336	307	29
17	glass-0-1-6_vs_2	10.3	9	192	175	17
18	ecoli-0-6-7_vs_5	10.0	6	220	200	20
19	vowel0	10.0	13	988	898	90
20	yeast-0-5-6-7-9_vs_4	9.4	8	528	477	51
21	ecoli-0-3-4-7_vs_5-6	9.3	7	257	232	25
22	ecoli-0-3-4-6_vs_5	9.3	7	205	185	20
23	glass-0-4_vs_5	9.2	9	92	83	9
24	ecoli-0-2-6-7_vs_3-5	9.2	7	224	202	22
25	ecoli-0-1_vs_2-3-5	9.2	7	244	220	24
26	ecoli-0-4-6_vs_5	9.2	6	203	183	20
27	yeast-0-2-5-6_vs_3-7-8-9	9.1	8	1,004	905	99
28	yeast-0-2-5-7-9_vs_3-6-8	9.1	8	1,004	905	99
29	yeast-0-3-5-9_vs_7-8	9.1	8	506	456	50
30	glass-0-1-5_vs_2	9.1	9	172	155	17
31	ecoli-0-2-3-4_vs_5	9.1	7	202	182	20
32	ecoli-0-6-7_vs_3-5	9.1	7	222	200	22
33	yeast-2_vs_4	9.1	8	514	463	51
34	ecoli-0-3-4_vs_5	9.0	7	200	180	20

The experimentation of this research is to demonstrate that our novel heuristic method can help improving the performance of classification model in various application areas with different imbalance ratios. So, all of 56 datasets are collected from 2 famous real-world dataset repositories, which are "KEEL" and "KDD-CUP". Then, we group them into two groups of imbalance ratio, that is, "low imbalance ratio" with a range of imbalance ratio from 9 to 20, and "high imbalance ratio" which imbalance ratio is over 20 and the maximum of imbalance ratio is 129. The Table I shows 34 datasets of "low imbalance ratio" and Table II show 24 datasets of "high imbalance ratio".

TABLE II: The 26 Datasets of "High Imbalance Ratio" Showing Numbers of Majority Class and Minority Class

Group High Imbalance Ratio						
No.	Dataset	IR	# Attr.	# Ins.	# Major	# Minor
1	abalone19	129.4	8	4,174	4,142	32
2	kddcup-rootkit-imap_vs_back	100.1	41	2,225	2,203	22
3	poker-8_vs_6	85.9	10	1,477	1,460	17
4	poker-8-9_vs_5	82.0	10	2,075	2,050	25
5	kr-vs-k-zero_vs_fifteen	80.2	6	2,193	2,166	27
6	kddcup-land_vs_satan	75.7	41	1,610	1,589	21
7	kddcup-buffer_overflow_vs_back	73.4	41	2,233	2,203	30
8	abalone-20_vs_8-9-10	72.7	8	1,916	1,890	26
9	winequality-red-3_vs_5	68.1	11	691	681	10
10	shuttle-2_vs_5	66.7	9	3,316	3,267	49
11	poker-8-9_vs_6	58.4	10	1,485	1,460	25
12	winequality-white-3-9_vs_5	58.3	11	1,482	1,457	25
13	kr-vs-k-zero_vs_eight	53.1	6	1,460	1,433	27
14	yeast6	41.4	8	1,484	1,449	35
15	ecoli-0-1-3-7_vs_2-6	39.1	7	281	274	7
16	yeast5	32.7	8	1,484	1,440	44
17	yeast-1-2-8-9_vs_7	30.6	8	947	917	30
18	yeast4	28.1	8	1,484	1,433	51
19	glass5	22.8	9	214	205	9
20	yeast-1-4-5-8_vs_7	22.1	8	693	663	30
21	yeast-2_vs_8	21.3	8	482	426	20
22	shuttle-c2-vs-c4	20.5	9	129	123	6

B. A Novel Heuristic Method

The novel heuristic method that we present in this paper is extracted from experience over 5 years of data mining expert engineers in the manufacturing field. The formula of this novel heuristic method is the relation between misclassification cost, imbalance ratio, and the constant e which is the "Euler's number" (~2.71828…). The computation of this heuristic is shown in equation 5.

$$\text{MCC} = \sqrt{\frac{\text{IR}^2}{e}} \tag{5}$$

where

MCC = misclassification cost or cost sensitive,
IR = imbalance ratio, and
e = Euler's number (constant number ~2.718…).

IR or imbalance ratio is defined by the calculation as shown in equation 6.

$$IR = \frac{\text{Number of majority class}}{\text{Number of minority class}} \tag{6}$$

We empirically validate this proposed method and have been found that it can improve classification performance in terms of the true positive rate in root cause analysis of computer's component manufacturing datasets with IR in the range of 4.1 to 1,245.7.

IV. Research Framework and Research Workflow

A. Research Framework

In this paper, we use 56 real-world datasets from serval areas such as medical/scientific experiment, wine quality, and many others. The minimum of imbalance ratio is 9 and the maximum is 129. These 56 datasets are classified into two groups: "low imbalance ratio" and "high imbalance ratio". The model that we use for classification in this paper is the state of art model in IBM SPSS Modeler, C5.0 model

(research framework is shown in Fig. 3). Then, we compare the model result between the traditional method and our novel heuristic method. The assumption of comparison in this paper focuses on 2 points:

1) The novel method should show better performance than the traditional method.
2) The high imbalance ratio group should show better of improvement rate than the low imbalance ratio group.

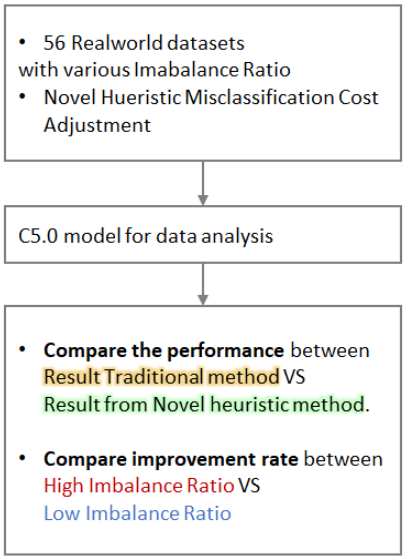

Fig. 3. Research framework.

B. Research Workflow

The research workflow of this research is shown in Fig. 4. Each of the 56 real-world datasets is used as input into the C5.0 model with 70% data instances for training the model and keep aside 30% of the rest for model validation. The same datasets are operated with two methods: "Traditional Method" and "Novel Heuristic Method Misclassification Cost Adjustment". After we run through this process we will obtain two classifiers from the two methods. Then the 30% of data that we set aside in earlier step will be used to test performance of classifiers. The final step is comparing the model performance in terms of AUC.

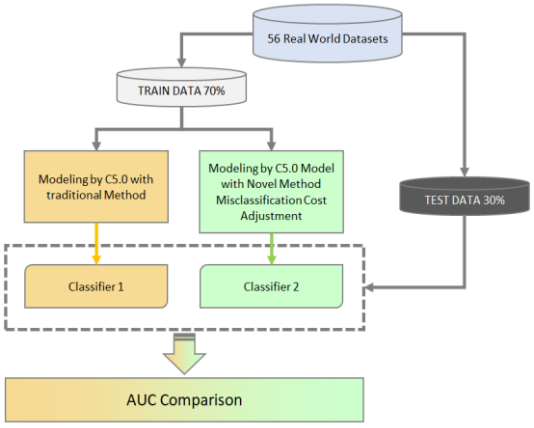

Fig. 4. Research workflow.

V. Experimentation and Results

This section is a demonstration of the experimentation results. The key point is a comparison between traditional method and novel heuristic method (or called proposed method). Table III is the experimentation results of "low imbalance ratio" group. There are 34 datasets in this group. The average imbalance ratio is 11.3 (minimum is 9 and maximum is 19.4), misclassification cost is averaged as 0.81.

In terms of AUC comparison, average AUC before adjusting misclassification cost (traditional methods) is 0.81 and after adjusting misclassification cost with the proposed method, AUC is 0.93. The proposed method shows the better AUC with the improvement rate of 18%. The top-3 of improvement rate are the dataset named "cleveland-0_ vs_4", "glass-0-1-6_vs_5" and "glass-0-1-6_vs_5" with the improvement rate of 97%, 82% and 57%, respectively.

The improvement rate is calculated by equation 7.

$$\text{Improvement Rate} = \frac{\text{Proposed AUC} - \text{Traditional AUC}}{\text{Traditional AUC}} \quad (7)$$

TABLE III: AUC Comparison between "Traditional Method" and "Propose Method" in Group "Low Imbalance Ratio"

				AUC : Area Under ROC Curve				
Group "Low Imbalance Ratio"				Traditional Method		Proposed Method		Improvement Rate
No.	Dataset	IR	MCC	Training	Testing	Training	Testing	
1	glass-0-1-6_vs_5	19.4	11.8	1.00	0.50	0.95	0.91	82%
2	abalone9-18	16.4	10.0	0.82	0.61	0.92	0.68	11%
3	page-blocks-1-3_vs_4	15.9	9.6	1.00	1.00	1.00	1.00	0%
4	ecoli4	15.8	9.6	0.88	0.74	0.97	0.88	18%
5	glass4	15.5	9.4	1.00	0.57	0.98	0.89	57%
6	yeast-1_vs_7	14.3	8.7	0.78	0.79	0.92	0.96	23%
7	shuttle-c0-vs-c4	13.9	8.4	1.00	1.00	1.00	1.00	0%
8	ecoli-0-1-4-6_vs_5	13.0	7.9	0.93	0.90	0.98	0.98	9%
9	cleveland-0_vs_4	12.6	7.7	0.88	0.46	0.97	0.90	97%
10	ecoli-0-1-4-7_vs_5-6	12.3	7.4	0.90	0.61	0.99	0.90	47%
11	glass2	11.6	7.0	0.97	0.92	0.96	0.92	0%
12	glass-0-1-4-6_vs_2	11.1	6.7	0.90	0.85	0.96	0.93	9%
13	ecoli-0-1_vs_5	11.0	6.7	0.99	0.75	0.97	0.93	24%
14	glass-0-6_vs_5	11.0	6.7	0.99	1.00	0.99	1.00	0%
15	led7digit-0-2-4-5-6-7-8-9_vs_1	11.0	6.7	0.96	0.94	0.97	0.94	0%
16	ecoli-0-1-4-7_vs_2-3-5-6	10.6	6.4	0.88	0.91	0.98	0.95	5%
17	glass-0-1-6_vs_2	10.3	6.2	0.50	0.50	0.94	0.79	57%
18	ecoli-0-6-7_vs_5	10.0	6.1	0.92	0.79	0.99	0.93	18%
19	vowel0	10.0	6.1	0.99	0.94	1.00	1.00	7%
20	yeast-0-5-6-7-9_vs_4	9.4	5.7	0.85	0.87	0.96	0.91	5%
21	ecoli-0-3-4-7_vs_5-6	9.3	5.6	0.84	0.90	0.90	1.00	11%
22	ecoli-0-3-4-6_vs_5	9.3	5.6	0.89	0.83	0.98	0.91	9%
23	glass-0-4_vs_5	9.2	5.6	0.99	1.00	0.99	1.00	0%
24	ecoli-0-2-6-7_vs_3-5	9.2	5.6	0.90	0.83	0.93	0.92	10%
25	ecoli-0-1_vs_2-3-5	9.2	5.6	0.99	0.77	0.99	0.90	16%
26	ecoli-0-4-6_vs_5	9.2	5.6	0.87	0.74	0.99	0.95	28%
27	yeast-0-2-5-6_vs_3-7-8-9	9.1	5.5	0.78	0.80	0.86	0.92	14%
28	yeast-0-2-5-7-9_vs_3-6-8	9.1	5.5	0.90	0.86	0.99	0.91	5%
29	yeast-0-3-5-9_vs_7-8	9.1	5.5	0.79	0.88	0.93	0.89	1%
30	glass-0-1-5_vs_2	9.1	5.5	0.84	0.71	0.95	0.92	30%
31	ecoli-0-2-3-4_vs_5	9.1	5.5	0.93	0.92	0.98	0.97	6%
32	ecoli-0-6-7_vs_3-5	9.1	5.5	0.85	1.00	0.98	1.00	0%
33	yeast-2_vs_4	9.1	5.5	1.00	0.96	0.98	0.98	3%
34	ecoli-0-3-4_vs_5	9.0	5.5	0.84	0.83	0.99	0.90	8%
AVG		**11.3**	**6.8**	0.90	**0.81**	0.97	**0.93**	**18%**

Table IV is the experimental results of "high imbalance ratio" showing comparative AUC performance between "Traditional Method" and "Proposed Method". In this groups, there are 22 datasets. The average value of imbalance ratio is 57.39 (minimum is 20.5 and maximum is 129). Average of misclassification cost adjustment is 34.8. AUC of traditional method is 0.74 compared to 0.90 of the proposed method. There are many datasets showing better performance in terms of AUC with high improvement rate.

The top-5 datasets are "yeast-1-4-5-8_vs_7", winequality-red-3_vs_5", "poker-8-9_vs_6", "poker-8_vs _6" and "winequality-white-3-9_vs_5". The improvement

rates are 77%, 71%, 67%, 65% and 63%, respectively. The average improvement rate is as high as 29%. It is a significant gap when compared to "low imbalance ratio" (which has an improvement rate of 18%).

TABLE IV: AUC COMPARISON BETWEEN "TRADITIONAL METHOD" AND "PROPOSE METHOD" IN GROUP "HIGH IMBALANCE RATIO"

				AUC : Area Under ROC Curve				
Group "High Imbalance Ratio"				Traditional Method		Proposed Method		Improvement Rate
No.	Dataset	IR	MCC	Training	Testing	Training	Testing	
1	abalone19	129.4	78.5	0.50	0.50	0.91	0.64	28%
2	kddcup-rootkit-imap_vs_back	100.1	60.7	1	1	1	1	0%
3	poker-8_vs_6	85.9	52.1	0.5	0.5	0.806	0.826	65%
4	poker-8-9_vs_5	82.0	49.7	0.5	0.5	0.849	0.718	44%
5	kr-vs-k-zero_vs_fifteen	80.2	48.7	0.932	0.801	0.965	0.964	20%
6	kddcup-land_vs_satan	75.7	45.9	1	1	1	1	0%
7	kddcup-buffer_overflow_vs_back	73.4	44.5	1	1	1	1	0%
8	abalone-20_vs_8-9-10	72.7	44.1	0.839	0.739	0.941	0.869	18%
9	winequality-red-3_vs_5	68.1	41.3	0.5	0.5	0.925	0.857	71%
10	shuttle-2_vs_5	66.7	40.4	1	1	1	1	0%
11	poker-8-9_vs_6	58.4	35.4	0.5	0.5	0.857	0.837	67%
12	winequality-white-3-9_vs_5	58.3	35.3	0.648	0.578	0.916	0.945	63%
13	kr-vs-k-zero_vs_eight	53.1	32.2	0.988	1	0.982	0.984	-2%
14	yeast6	41.4	25.1	0.92	0.75	0.90	0.96	28%
15	ecoli-0-1-3-7_vs_2-6	39.1	23.7	0.80	0.99	0.98	0.99	0%
16	yeast5	32.7	19.9	0.99	0.91	0.99	0.98	8%
17	yeast-1-2-8-9_vs_7	30.6	18.5	0.66	0.49	0.93	0.70	43%
18	yeast4	28.1	17.0	0.86	0.80	0.96	0.87	8%
19	glass5	22.8	13.8	1.00	0.63	0.94	0.98	57%
20	yeast-1-4-5-8_vs_7	22.1	13.4	0.50	0.50	0.89	0.89	77%
21	yeast-2_vs_8	21.3	12.9	0.50	0.50	0.79	0.75	50%
22	shuttle-c2-vs-c4	20.5	12.4	1.00	1.00	1.00	1.00	0%
AVG		57.39	34.81	0.78	0.74	0.93	0.90	29%

VI. CONCLUSION

In this paper, we presented the novel heuristic method to compute proper cost-sensitive value for classifying imbalanced data that have high imbalance ratio between the tremendous majority class as compared to the tiny minority class. The experimentation have been performed on the 56 real-world datasets to assess the improvement rate of AUC when compared to the traditional classification method. These datasets are from various domains and various imbalance ratios. The key proposals of this paper are based on the two assumptions:

- Novel method can improve the model performance when compared to traditional classification method.
- High imbalance ratio should show the better improvement rate than low imbalance ratio.

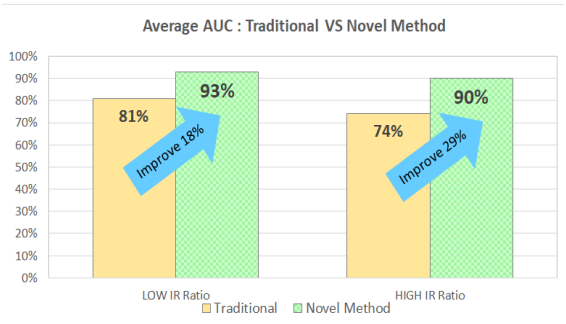

Fig. 5. Summary graph showing overall AUC comparisons improvement rate between "traditional method" and "propose method".

It turns out that the experimental results confirm our assumptions. From overall data, we can see the improvement rate at the satisfying level. For the 34 datasets of low imbalance group, with the imbalance ratio ranging from 9 to 20, the improvement rate is about 18%. For the 22 datasets of high imbalance ratio (with imbalance ratio over 20), the improvement rate is 29% on average. A graph of overall AUC comparisons is shown in Fig. 5. From this result, we can conclude that our novel heuristic method is suitable for classifying data with high imbalance ratio.

VII. RECOMMENDATION

On standard datasets obtained from the worldwide repositories, we observe that imbalance ratios in these data are not so high (11.3 to 57.39 on average). This is unlike real production data of manufacturing fields in which the imbalance ratio can be as high as 1: 1,000 or over. Based on the experimental results that reveal significant classification improvement when the imbalance ratio is very high, we thus expect that the proposed novel heuristic method can show clearly the improvement over traditional classification when the imbalance ratio of manufacturing data is in extreme level.

In our further research, we plan to use this method in misclassification cost adjustment with data in other fields that have extremely high level of imbalance ratio. Moreover, the multiclass target classification is also the challenging area that we would like to tackle with this method.

ACKNOWLEDGMENT

This research work has been supported by grants from the National Research Council of Thailand (NRCT). The first author has been support by the scholarship from Suranaree University of Technology. All three authors are researchers of the Data and Knowledge Engineering Research Unit that has been fully supported by research grant from Suranaree University of Technology.

REFERENCES

[1] W.-L. Chao, J.-Z. Liu, and J.-J. Ding, "Facial age estimation based on label-sensitive learning and age-oriented regression," *Pattern Recognit.*, vol. 46, no. 3, pp. 628–641, 2013.
[2] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, no. 2–3, pp. 195–215, 1998.
[3] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "Adaptive ROC-based ensembles of HMMs applied to anomaly detection," *Pattern Recognit.*, vol. 45, no. 1, pp. 208–230, 2012.
[4] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 291–316, 1997.
[5] L. Pelayo and S. Dick, "Applying novel resampling strategies to software defect prediction," in *Fuzzy Information Processing Society, 2007. NAFIPS'07. Annual Meeting of the North American*, 2007, pp. 69–72.
[6] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognit.*, vol. 45, no. 1, pp. 346–362, 2012.
[7] G. M. Weiss, "Mining with rarity: A unifying framework," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.
[8] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.
[9] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, no. 9, pp. 1263–1284, 2008.
[10] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: A case study involving information extraction," in *Proc. Workshop on Learning from Imbalanced Datasets*, 2003, vol. 126.
[11] W. Liu and S. Chawla, "Class confidence weighted knn algorithms for imbalanced data sets," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2011, pp. 345–356.
[12] F. Provost, "Machine learning from imbalanced data sets 101," in *Proc. the AAAI workshop on imbalanced data sets*, 2000, pp. 1–3.
[13] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, 2007.

[14] X.Y. Liu, J. Wu, and Z.H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst. Man, Cybern. Part B*, vol. 39, no. 2, pp. 539–550, 2009.
[15] R. Barandelaa, J. S. Sanchezb, and V. Garcia, "Strategies for learning in class imbalance problems," 2003.
[16] M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognit.*, vol. 45, no. 10, pp. 3738–3750, 2012.
[17] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[18] S. Garcia and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evol. Comput.*, vol. 17, no. 3, pp. 275–306, 2009.
[19] G. M. Weiss, K. McCarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error cost," *DMIN*, vol. 7, pp. 35–41, 2007.
[20] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, 2006.
[21] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "A comparative study of data sampling and cost sensitive learning," in *Proc. IEEE International Conference on Data Mining Workshops, 2008*, 2008, pp. 46–52.
[22] P. Su, W. Mao, and D. Zeng, "An empirical study of cost-sensitive learning in cultural modeling," *Inf. Syst. E-bus. Manag.*, vol. 11, no. 3, pp. 437–455, 2013.
[23] S. Lomax and S. Vadera, "A survey of cost-sensitive decision tree induction algorithms," *ACM Comput. Surv.*, vol. 45, no. 2, p. 16, 2013.
[24] J. R. Quinlan, "Discovering rules by induction from large collections of examples," *Expert Syst. Micro Electron. Age*, 1979.
[25] J. R. Quinlan, "Learning efficient classification procedures and their application to chess end games," *Machine Learning*, Elsevier, vol. I, pp. 463–482, 1983.
[26] L. A. Breslow and D. W. Aha, "Simplifying decision trees: A survey," *Knowl. Eng. Rev.*, vol. 12, no. 1, pp. 1–40, 1997.
[27] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, Elsevier, 2014.
[28] S. Bertolini, A. Maoli, G. Rauch, and M. Giacomini, "Entropy-driven decision tree building for decision support in gastroenterology," *Stud Heal. Technol Inf.*, vol. 186, pp. 93–97, 2013.
[29] T. Wendler and S. Gröttrup, *Data mining with SPSS Modeler: Theory, Exercises and Solutions*, Springer, 2016.
[30] 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, 2004.
[31] W. Lu, Z. Li, and J. Chu, "Adaptive ensemble undersampling-boost: A novel learning framework for imbalanced data," *J. Syst. Softw.*, vol. 132, pp. 272–282, 2017.
[32] W. C. Lin, C. F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Inf. Sci. (Ny).*, vol. 409, pp. 17–26, 2017.
[33] L. Peng, H. Zhang, B. Yang, and Y. Chen, "A new approach for imbalanced data classification based on data gravitation," *Inf. Sci. (Ny).*, vol. 288, pp. 347–373, 2014.
[34] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.
[35] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Glob. Ecol. Biogeogr.*, vol. 17, no. 2, pp. 145–151, 2008.
[36] L. Sun, J. Wang, and J. Wei, "AVC: Selecting discriminative features on basis of AUC by maximizing variable complementarity," *BMC Bioinformatics*, vol. 18, no. 3, p. 50, 2017.

Anusara Hirunyawanakul is a Ph.D. student, School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. She received her B.E. and M.E. in computer engineering from Suranaree University of Technology, Thailand, in 2006 and 2014. Her research of interest includes Data Mining Applications, Machine Learning, and Artificial Intelligence in Manufacturing.

Nittaya Kerdprasop is an associate professor and the head of Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. She received her B.S. in radiation techniques from Mahidol University, Thailand, in 1985, M.S. in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. Her research of interest includes Data Mining, Artificial Intelligence, Logic and Constraint Programming.

Kittisak Kerdprasop is an associate professor at the School of Computer Engineering, Chair of the School, and the head of Knowledge Engineering Research Unit, SUT. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, MS in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes Machine Learning and Artificial Intelligence.