CHAPTER 4
EXPERIMENTAL RESULTS AND ANALYSIS

In order to investigate suggested oversampling techniques, i.e. triangular SMOTE, relocating safe-level SMOTE and adaptive neighbor SMOTE, may improve the classification performance in the class imbalance problem, empirical experiments are conducted and the comparison and analysis of their results are presented in this chapter. In the first part of the chapter, the description of benchmark datasets and the experimental setting and the related statistical test are described. Then, the results of oversampling techniques against other existing oversampling techniques are shown. The number of cases each technique achieves the best and top three highest ranks for F-measure, geometric mean and adjusted g-mean is counted. The second part of chapter covers the statistical tests to show the significant improvement of these algorithms.

4.1 Datasets and experimental settings

4.1.1 The description of benchmark datasets

For this dissertation, experiments are performed on 9 datasets from UCI repository [91]; ecoli, glass, letter recognition, haberman, LandSat(satimage), segmentation, yeast, optdigits and vehicle, and 5 datasets from PROMISE repository [92]; cm1, jm1, kc1, kc2 and pc1. These datasets are numerical and contain no missing values. Moreover, they are either binary class datasets with an unequal class distribution or multiple class datasets which can be transformed into the binary class dataset with an unequal class distribution by selecting one class as positive and treating others as negative. The description about the number of instances, the number of attributes, the number of positive instances and the percentage of positive instances are shown in table 6 below.

Table 6: The description of datasets used in the experiments.

Name	Instances	Attributes	Positive instances	% of positive instances
cm1	498	21	49	10.91
Ecoli	336	8	20	5.95
Glass	214	11	76	35.51
Haberman	306	4	81	26.47
Letter (H)[1]	20,000	17	734	3.67
jm1	10,880	21	2,103	23.96
kc1	2,109	21	326	18.28
kc2	522	21	107	25.78
Optdigits (0)[1]	5,620	64	554	10.94
pc1	1,109	21	77	7.46
Satimage (4)[1]	6,435	37	626	9.73
Segment (WIN)[1]	2,310	20	330	14.29
vehicle	846	18	218	34.71
Yeast (ME3)[1]	1,484	9	163	10.98

4.1.2 Experimental settings

The experiments are conducted for five classifiers; decision tree (C4.5) [32], naïve Bayes classifier [8], multilayer perceptron [9], support vector machine [10] with the linear square kernel and *k*-nearest neighbor [11] (with $k = 3$). These classifiers are standard algorithms in classification and they are included in most data mining softwares. Different classifiers are used as candidates for various framework. The oversampling techniques introduced in chapter 3 are expected to perform well in all classifiers. The setting for a classifier uses its default setting from the data mining software, KNIME [34]. The performance is evaluated through the train-test evaluation. The training set is stratified sampled from 70% of each benchmark dataset and 30%

[1] For multiclass datasets, the name of class shown in the parenthesis next to the name of dataset is the one used as positive in this experiment, all other classes of that dataset are used as negative.

is used as the test set. Each dataset is sampled 50 times giving 50 different pairs of training-test datasets. The training set from each sampling technique is used to generate synthetic minority instances through oversampling techniques in R programming environment [33]. For RSLS and ANS which have an additional process of minority outcast handling, the minority outcast detection is performed.

The number of minority outcasts depends on the value of *c*. In order to find an appropriate value of *c*, an experiment is conducted on each benchmark dataset by running *c*-nearest neighbor on it. The value of *c* is varied from 1 to 20 and the number of minority outcasts for each value of *c* in each dataset is counted. The percentage of outcasts to a total number of positive instances in all 14 datasets used for this experiment is plotted. The darker line is the plot of average percentages. The graph is presented in figure 20.

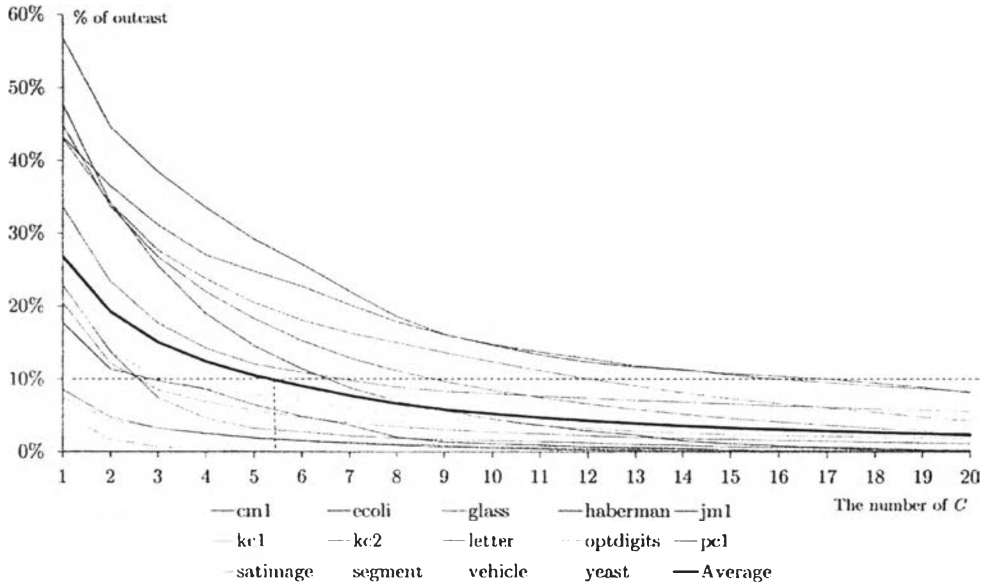


Figure 20: The graph showing the percentage of outcast instances in each dataset when the value of *c* is varied.

In this dissertation, 10% of positive instances are set as the prefer number of outcasts. The graph leads to the value of *c* equal to 5. It coincides with the setting of *c* in safe-level SMOTE [28] which also equals to 5. Therefore, the value of *c* as 5 is applied for each oversampling technique which contains the *c*-nearest neighbor process in this dissertation.

After positive instances which are not minority outcast instances in the training set are sent to each oversampling technique to synthetize instances, the resulting synthetic instances are added into the original training set along with a set of minority outcast instances which is extracted as a part of RSLS and ANS algorithm. This synthetic balanced training set from each oversampling technique and a set of outcast instances are sent into KNIME to perform the classification with five classifiers.

In RSLS and ANS, a set of minority outcasts is combined with a set of negative instances from each training set to train a 1-nearest neighbor model as the process of minority outcast handling. This model is applied on unknown instances in the test set after the classification stage as mentioned in chapter 3. The evaluation of classification on various performance measures are also performed in KNIME [34]. Performance measures used for evaluating the performance are F-measure which takes account of both recall and precision, geometric mean and adjusted g-mean [31] which consider the prediction rate in both classes simultaneously. The results in term of geometric mean and adjusted g-mean are reported in the appendix.

The diagram of the experimental process in each experimental round in one case of a classifier and a benchmark dataset is shown in figure 21.

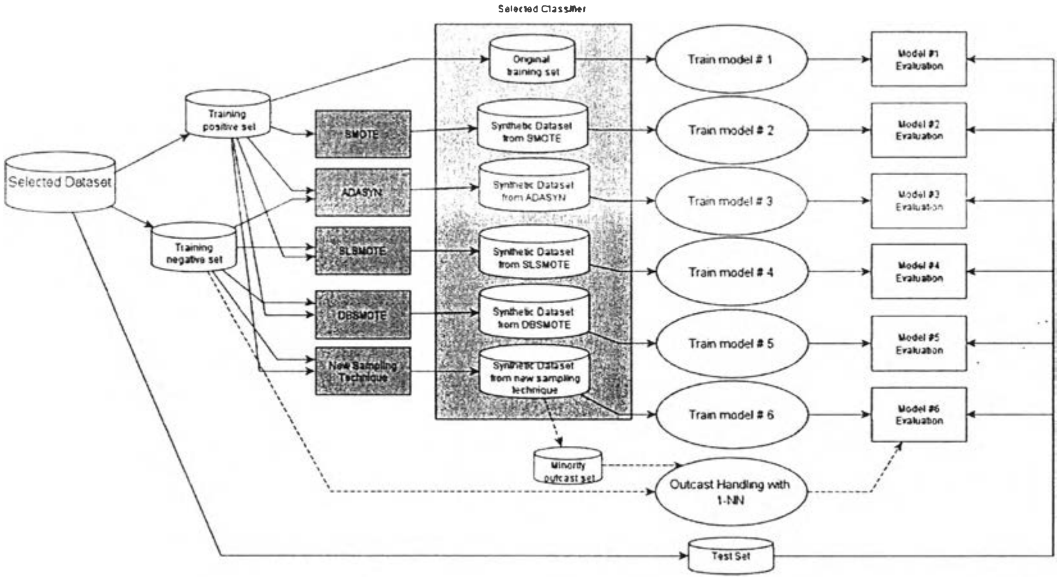


Figure 21: The diagram of the experimental process in each round of train-test sampling

From this setting, the total number of cases of original dataset and classifier is 70 (14 datasets x 5 classifiers). There are 50 rounds of train-test sampling in each case.

4.1.3 Wilcoxon signed-rank test

The Wilcoxon signed-rank test [30] is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to evaluate whether their population mean ranks differ. It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed. The Wilcoxon signed-rank test assumptions are

1. Data are paired and come from the same population.

2. Each pair is chosen randomly and independent and the order of pair has no significance

3. The data does not require being normally distributed but they are measured on an ordinal, interval, or ratio scale.

4. The distribution of the differences is symmetric around the median.

The null hypothesis for the two-tailed Wilcoxon signed-rank test is usually that the median difference between pairs of observations is zero. Note that this is different from the null hypothesis of the paired t-test, which is that the mean difference between pairs is zero, or the null hypothesis of the sign test, which is that the numbers of differences in each direction are equal. The null hypotheses for the two-tailed test and each tail of one-tailed test and their counterpart alternative hypotheses are shown in the following table.

Table 7: The null and alternative hypotheses in each type of Wilcoxon signed-rank test

Two-tailed Test	One-tailed test in the lower tail	One-tailed test in the upper tail
$H_0 : M_{diff} = 0$	$H_0 : M_{diff} \geq 0$	$H_0 : M_{diff} \leq 0$
$H_1 : M_{diff} \neq 0$	$H_1 : M_{diff} < 0$	$H_1 : M_{diff} > 0$

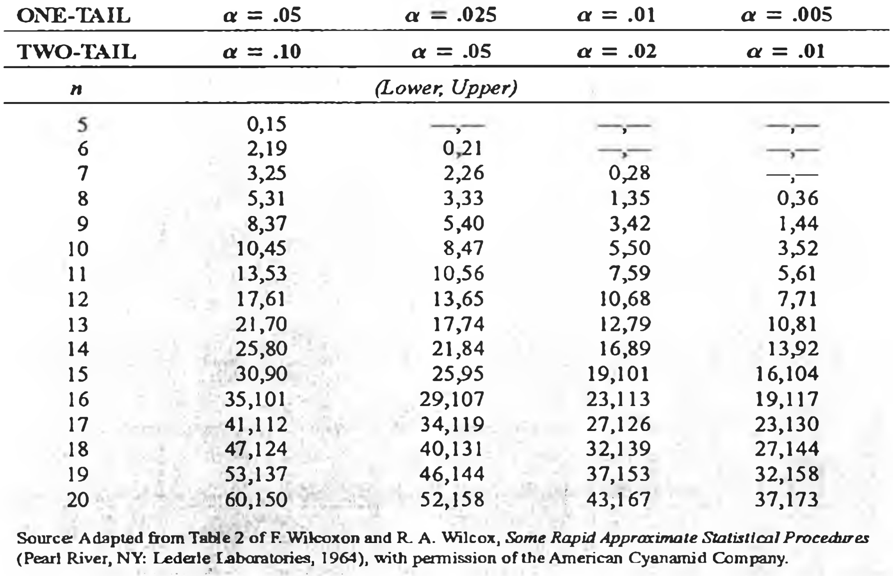
The basic procedure of the Wilcoxon signed-rank test is (1) setting the significant level α, (2) extracting the sample, (3) computing the value of the Wilcoxon test statistic W and comparing it with the critical upper bound and lower bound values which depend on whether the test is two-tailed or one-tailed. If the computed W test statistic equals to or is greater than the upper critical value (for two-tailed test and one-tailed test in the upper tail) or equals to or less than the lower critical value (for two-tailed test and one-tailed test in the lower tail), the null hypothesis is rejected. The process to compute the value of the Wilcoxon test statistic *W* is given below.

Computing the Wilcoxon signed-ranks

1. For each item in a sample of *n* items, compute a difference score, D_i, between the two paired values.
2. Neglect the + and – signs and list the set of *n* absolute differences, $|D_i|$.
3. Omit any absolute difference score of zero from further analysis, thereby yielding a set of *n*' nonzero absolute difference scores, where $n' \leq n$. After values with absolute difference scores of zero are removed, reset *n*' to be the actual sample size.
4. Assign ranks from 1 to *n*' to each of the $|D_i|$ such that the smallest absolute difference score gets rank 1 and the largest score gets rank *n*'. If two or more $|D_i|$ are equal, assign each of them the mean of the ranks they would have been assigned individually.
5. Reassign the symbol + or - to each of the *n*' ranks, R_i, depending on whether D_i was originally positive or negative.
6. Compute the Wilcoxon test statistic, *W*, as the sum of the positive ranks

$$W = \sum_{i=1}^{n'} R_i^{(+)}$$

or $n' \leq 20$, the critical upper bound and lower bound values can be looked up from a given table in figure 22.



ONE-TAIL	α = .05	α = .025	α = .01	α = .005
TWO-TAIL	α = .10	α = .05	α = .02	α = .01
n		*(Lower, Upper)*		
5	0,15	—,—	—,—	—,—
6	2,19	0,21	—,—	—,—
7	3,25	2,26	0,28	—,—
8	5,31	3,33	1,35	0,36
9	8,37	5,40	3,42	1,44
10	10,45	8,47	5,50	3,52
11	13,53	10,56	7,59	5,61
12	17,61	13,65	10,68	7,71
13	21,70	17,74	12,79	10,81
14	25,80	21,84	16,89	13,92
15	30,90	25,95	19,101	16,104
16	35,101	29,107	23,113	19,117
17	41,112	34,119	27,126	23,130
18	47,124	40,131	32,139	27,144
19	53,137	46,144	37,153	32,158
20	60,150	52,158	43,167	37,173

Source: Adapted from Table 2 of F. Wilcoxon and R. A. Wilcox, *Some Rapid Approximate Statistical Procedures* (Pearl River, NY: Lederle Laboratories, 1964), with permission of the American Cyanamid Company.

Figure 22: The table of the critical upper and lower bound values of W when *n* is no more than 20.

For samples with *n*' > 20, the test statistic *W* value is normally distributed with the mean μ_W and standard deviation σ_W. The mean of the test statistic *W* is

$$\mu_W = \frac{n'(n'+1)}{4}$$

And the standard deviation of the test statistic *W* is

$$\sigma_W = \sqrt{\frac{n'(n'+1)(2n'+1)}{24}}$$

Then, the large-sample approximation formula with calculated *W* when *n*' > 20 is achieved as

$$Z_{STAT} = \frac{W - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}}$$

If the computed Z_{STAT} falls in the critical region (For α = 0.05, the confidence interval is between ±1.96 which means the value outside this range is the critical value), the null hypothesis is rejected.

For R programming, the function **wilcox.test()** is used for performing the test. The function provides the median difference and the p-value. If this p-value is less than α, the null hypothesis is rejected. It can perform either one-tailed test or two-tailed test by assigned the desired alternate hypothesis in one of its arguments.

4.2 The result analysis

4.2.1 Triangular minority oversampling technique

In this section, the experiment on triangular minority oversampling technique (TMOT) which is introduced in chapter 3 is performed to compare its performance on generating balanced dataset that can train classifiers to classify minority class effectively against the original imbalanced dataset and the balanced dataset from SMOTE [25]. The average of F-measure results from each oversampling technique in 50 rounds of experiment is compared between datasets and classifiers. Figure 23-Figure 27 are plotted among the result from the original imbalanced dataset (ORIG), the balanced dataset from SMOTE and TMOT.

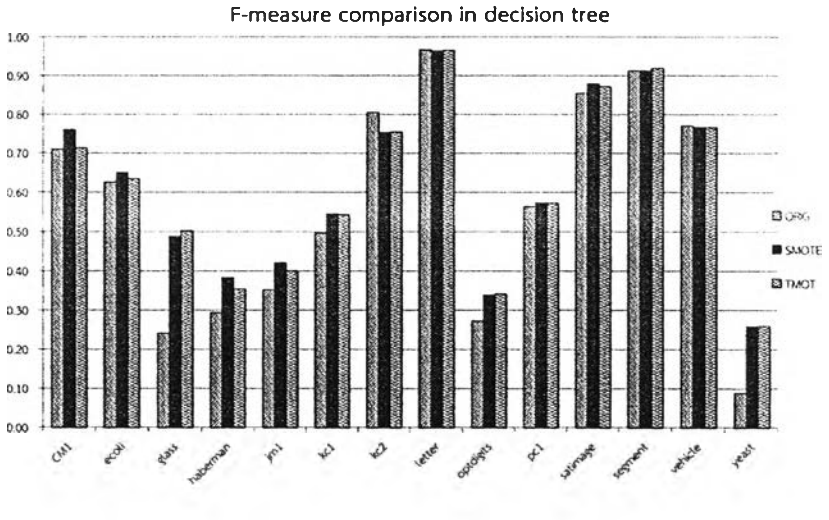


Figure 23: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a decision tree as a classifier

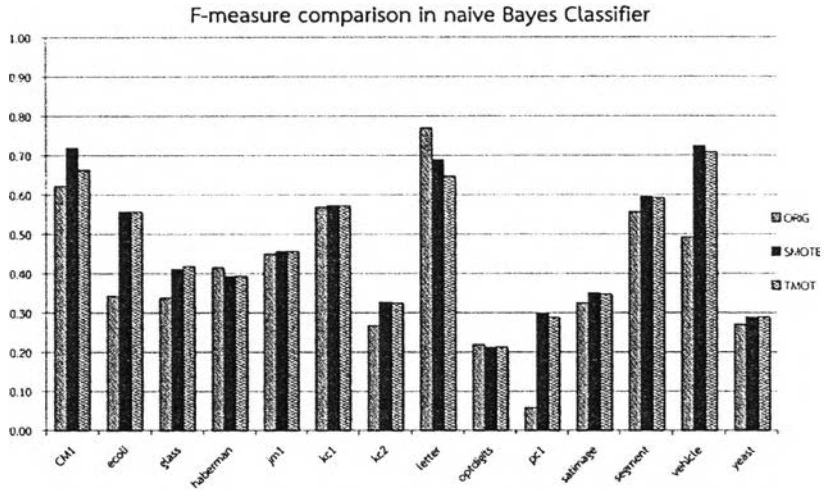


Figure 24: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a naïve Bayes classifier as a classifier

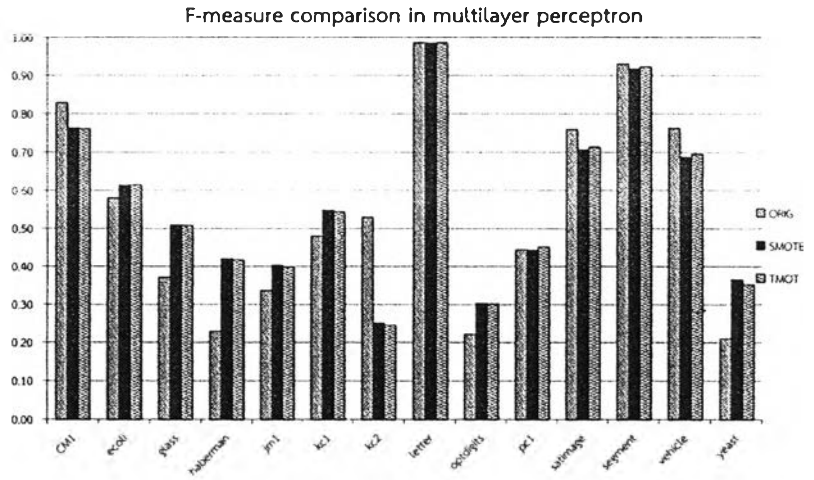


Figure 25: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a multilayer perceptron as a classifier

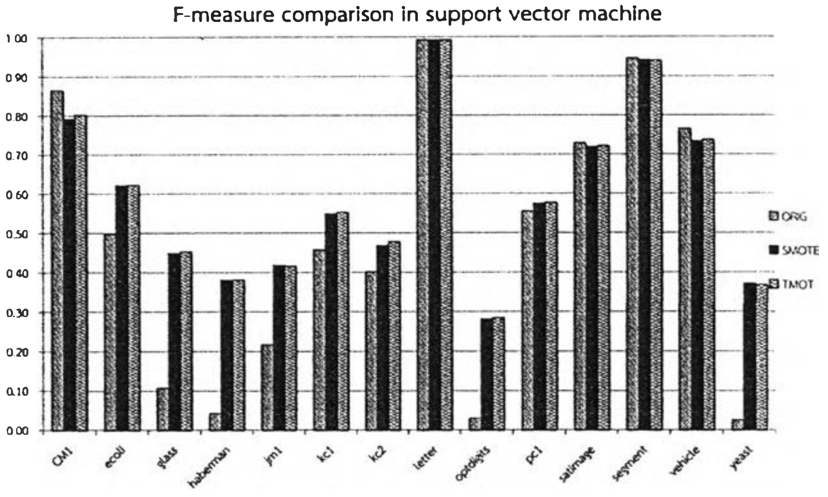


Figure 26: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a support vector machine as a classifier

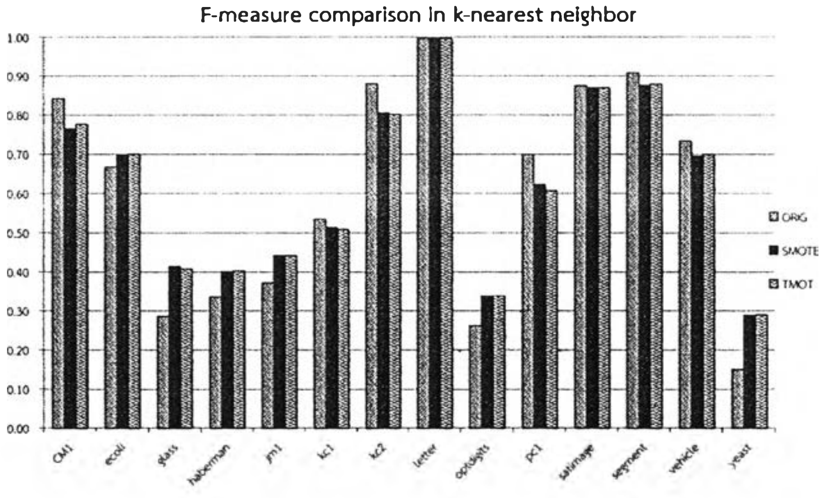


Figure 27: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a *k*-nearest neighbor as a classifier

Figure 23 to figure 27 show the bar charts of the mean comparison of F-measure each oversampling technique achieves in each classifier where ORIG refers to the result of a classifier performing on an original imbalanced dataset, SMOTE refers to the result of a classifier performing on the balanced dataset applying synthetic minority oversampling technique and TMOT refers to the result of a classifier performing on the balanced dataset applying triangular minority oversampling technique. There are 21 cases which TMOT achieves the average F-measure higher than both SMOTE and ORIG. Most cases occur when support vector machine is chosen as a classifier. There are another 26 cases which TMOT can defeat ORIG but still has the lower average F-measure than SMOTE. It achieves a higher average F-measure than SMOTE but its average is lower than one from ORIG in 14 cases. TMOT has a lower average F-measure than both algorithms in only 9 cases.

In order to clarify whether TMOT is an effective oversampling technique comparing with SMOTE, the Wilcoxon signed-rank test is performed to test whether the difference between the F-measure from two algorithms are significant. The null hypothesis of the test is set as the median of difference is less than or equal to zero, so if a p-value of test is less than 0.05, then the alternative hypothesis which is the median of difference between the controlled algorithm (TMOT in this case) and the other compared algorithm are positive. Using the results from every round of experiments regardless of various classifiers and datasets through R programming environment [33], the test results are shown in table 8.

Table 8: The Wilcoxon signed-rank test on F-measures from TMOT against ones from ORIG and SMOTE

TMOT	Median of Difference	p-value	The number of nonzero difference pairs	sum of positive rank
against ORIG	0.0228	0.0000	3371	3919117
against SMOTE	-0.0005	0.9997	2977	2055314

It could be seen that the median of difference of F-measure from TMOT and F-measure from ORIG is significantly positive. It implies that using TMOT provides the better F-measure than using the original imbalanced dataset to build the classifier. However, the result comparing with SMOTE is not significantly different. The median of their difference is less than zero and the p-value is higher than 0.05. This means SMOTE and TMOT are not different in term of the performance. Note that TMOT uses more arithmetic calculation for creating a new synthetic instance but it yields the same result as SMOTE, It can be concluded that SMOTE is a more preferable oversampling technique in this regard.

4.2.2 Relocating safe-level SMOTE

In this section, the performance of relocating safe-level SMOTE (RSLS) is compared with other predecessor sampling techniques. As the extension from safe-level SMOTE [28], it is expected that RSLS should provide the better accuracy performance than one from safe-level SMOTE. In order to measure the performance, F-measure value of each classification via datasets from each oversampling techniques is focused. In each round of experiments, there are 6 different training sets; the original imbalanced training set without performing any sampling techniques (ORIG), the balanced training set containing synthetic instances from synthetic minority oversampling technique (SMOTE) [25], the balanced training set containing synthetic instances from adaptive synthetic sampling (ADASYN) [26], the balanced training set containing synthetic instances from safe-level SMOTE (SLS) [28], the balanced training set containing synthetic instances from density-based synthetic minority oversampling technique (DBSMOTE) [29] and the balanced training set containing synthetic instances from relocating safe-level SMOTE (RSLS) to build the classification model with each designated classifier. The resulting models from these training sets with respect to each classifier and dataset are evaluated using the same test set. For RSLS, minority outcast handling process is also performed.

The average F-measure values from 50 rounds of experiments in each classifier and dataset are shown and ranked in table 24 in appendix 1. In table 24, the dark gray shade highlights the F-measure value which is the highest value among F-measure from all oversampling techniques from the same classifier and dataset, while the light gray shade highlights the F-measure value which is one of the top three of oversampling techniques from the same classifier and dataset. The number of cases each technique can achieve the best F-measure is summarized as the bar chart in figure 28.

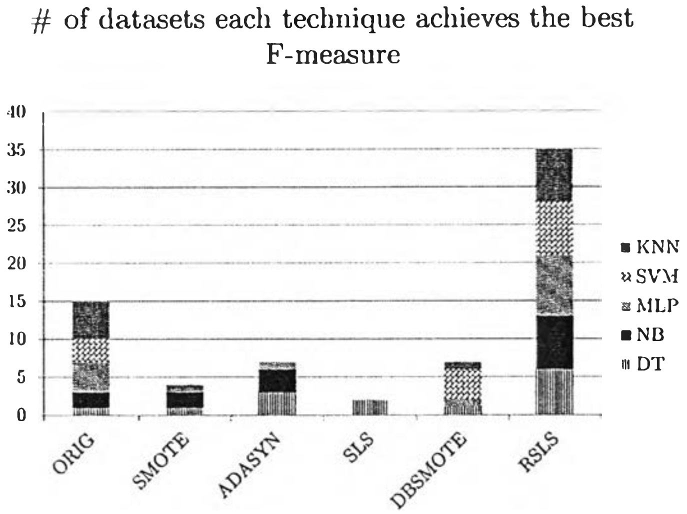


Figure 28: The bar chart of the number of datasets each oversampling technique achieves the best F-measure

The comparison of F-measure presented in table 24 and summarized in table 10 shows that RSLS provides relatively good performance on F-measure generally as it achieves top three F-measure from 60 out of 70 cases and the highest value from 35 cases after compared with other 5 oversampling techniques. From these 35 cases, 6 of them are achieved when C4.5 is used as the classifier. 7 of them come from naïve Bayes classifier. RSLS achieves the best F-measure value when multilayer perceptron is a classifier in 8 datasets. For both support vector machine and k-nearest neighbor, RSLS get the best F-measure in 7 datasets. The number of datasets RSLS can achieve the best F-measure is distributed nearly equally in each classifier showing that RSLS does not bias to one classifier. The list of datasets RSLS provide the best, second best and third best F-measure is shown in table 9 and the number of cases each oversampling technique achieves the best, second best and third best F-measure is shown in table 10.

Table 9: The list of dataset names which RSLS achieves the best, second best and third best F-measure in each classifier.

Classifier	Datasets in 1^{st}	Datasets in 2^{nd}	Datasets in 3^{rd}	The total number
Decision tree (C4.5)	ecoli, haberman, jm1, optdigits, pc1, vehicle	cm1, kc1	glass, kc2, yeast	11
Naïve Bayes classifier	ecoli, jm1, kc1, kc2, optdigits, satimage, yeast	glass, haberman, letter, segment	cm1, pc1, vehicle	14
Multilayer perceptron	cm1, ecoli, haberman, jm1, kc1, letter, optdigits, yeast		glass, pc1, segment, vehicle	12
Support vector machine	glass, haberman, jm1, kc1, letter, optdigits, yeast	ecoli, pc1,	cm1, kc2, satimage, segment, vehicle	14
K-nearest neighbor	ecoli, haberman, jm1, kc1, optdigits, vehicle, yeast	glass	cm1	9

Table 10: The number of cases each technique achieves the average F-measure in the ranking 1^{st} -3^{rd}

# of cases as	ORIG	SMOTE	ADASYN	SLS	DBSMOTE	RSLS
1^{st}	15	4	7	2	7	35
2^{nd}	5	15	9	22	10	9
3^{rd}	3	18	7	19	7	16
Total in 1^{st} -3^{rd}	23	37	23	43	24	60

The Wilcoxon signed-rank test is performed to verify the difference of F-measure from RSLS against other oversampling techniques. First, every experimental result from RSLS is used to compare pairwise with F-measure from each oversampling technique. The null hypothesis of the test is set as the median of difference is less or equal than zero, so if a p-value of test is less than 0.05, then the alternative hypothesis which is the median of difference between the controlled oversampling technique (RSLS in this case) and other compared oversampling techniques are positive. The results of these statistical tests against each oversampling technique are shown in table 11.

Table 11: The Wilcoxon signed-rank of the difference of F-measure from RSLS against other sampling techniques

RSLS against	The median of difference	p-value
ORIG	0.0441	5.5100×10^{-167}
SMOTE	0.0170	8.9000×10^{-110}
ADASYN	0.0255	1.9400×10^{-161}
SLS	0.0117	3.8300×10^{-72}
DBSMOTE	0.0272	4.4100×10^{-129}

In table 11, it shows that the p-value from the Wilcoxon signed-rank test is lower than 0.05 with all five techniques. By these p-values, the null hypothesis for each comparison is rejected. Consequently, the alternate hypothesis which is the median of difference is positive will be accepted. This result suggests that every difference of F-measure between RSLS and other oversampling techniques are significantly positive.

In table 12, the results are separated based on classifiers and performed the Wilcoxon signed-rank test in order to see whether there are significant difference of F-measure in each classifier. The result in the table shows that there is significantly positive difference of F-measure when comparing RSLS with other oversampling techniques in every classifier since every p-value in each test is less than 0.05.

Table 12: The Wilcoxon signed-rank of the difference of F-measure from RSLS against other sampling techniques in each classifier

Classifier	RSLS against	Median of Difference	p-value
DT	ORIG	0.0339	2.4800×10^{-38}
	SMOTE	0.0111	1.0500×10^{-09}
	ADASYN	0.0108	3.9200×10^{-09}
	SLS	0.0120	1.7300×10^{-13}
	DBSMOTE	0.0298	1.1100×10^{-30}
NB	ORIG	0.0448	4.6600×10^{-54}
	SMOTE	0.0152	5.2000×10^{-20}
	ADASYN	0.0254	1.3500×10^{-32}
	SLS	0.0186	1.3800×10^{-48}
	DBSMOTE	0.0847	2.0900×10^{-89}
MLP	ORIG	0.0290	1.1200×10^{-13}
	SMOTE	0.0175	1.6700×10^{-16}
	ADASYN	0.0288	1.1900×10^{-36}
	SLS	0.0090	1.9000×10^{-07}
	DBSMOTE	0.0177	2.0000×10^{-10}
SVM	ORIG	0.1124	1.9000×10^{-70}
	SMOTE	0.0265	1.0500×10^{-60}
	ADASYN	0.0363	4.0000×10^{-70}
	SLS	0.0083	8.4700×10^{-31}
	DBSMOTE	0.0059	1.8200×10^{-07}
KNN	ORIG	0.0151	6.1300×10^{-15}
	SMOTE	0.0145	1.3200×10^{-26}
	ADASYN	0.0229	2.5300×10^{-41}
	SLS	0.0036	1.9300×10^{-05}
	DBSMOTE	0.0081	5.0300×10^{-12}

4.2.3 Adaptive neighbors SMOTE

Similar with the comparison setting with RSLS, the average F-measure values from 50 rounds of experiments respect to each oversampling technique under the same classifier and benchmark dataset are compared. Each technique generates synthetic instances which are added into the same original imbalanced dataset, making it the balanced dataset. These resulting datasets are used to train classifier. Resulting classifiers are evaluated with the same test set in each round. Additionally, the minority outcast handling process is applied to improve the classification result of ANS model. The results from ANS models; the one with minority outcast handling (labeled as ANS2) and the one without minority outcast handling (labeled as ANS1) are both collected for comparison. The average F-measure value of ANS1 and other 5 oversampling techniques are ranked. Then, the number of datasets each technique provides the best F-measure and provides the top three F-measure of each case is counted and reported as bar charts in figure 29 to figure 30. The similar comparison is presented with the result of ANS2 in figure 31 to figure 32.

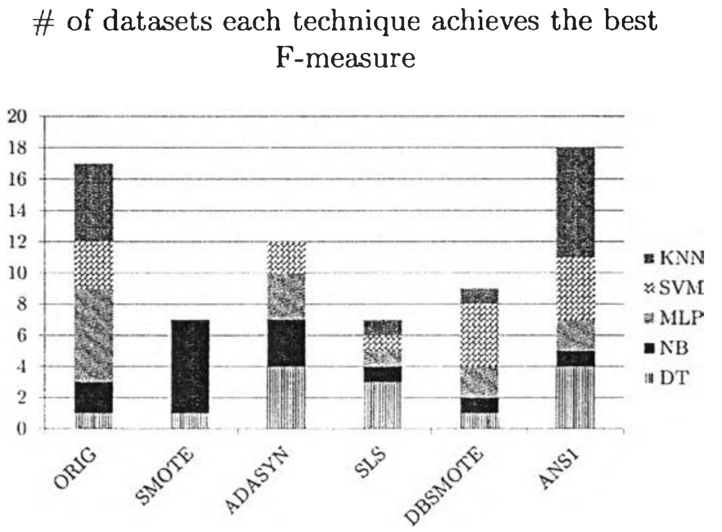


Figure 29: The bar chart of the number of datasets which ANS1 and each oversampling technique achieves the best F-measure

Figure 29 shows the bar chart which counts the number of datasets which ANS without outcast handling (ANS1) and other existing oversampling techniques can provide the best F-measure. It shows that ANS1 achieves the best F-measure in 18 cases out of total 70 cases which is the most among these 6 techniques. (Original wins 17, SMOTE wins 7, ADASYN wins 12, safe-level SMOTE wins 7 and DBSMOTE wins 9). ANS1 is the technique with the highest number of datasets with best F-measure in 3 classifiers and has the highest total number of datasets. The datasets which ANS1 has the best, second best and third best F-measures in each classifier are shown in table 13.

If achieving the top three F-measure is used to indicate the consistency of performance, it is presented as the bar chart in figure 30. The bar chart shows that the number of cases which ANS without minority outcast handling achieves the top F-measure is 46 which is 65% of the total number of cases. However, ANS1 is not the oversampling technique which has the highest number of cases on achieving the top three. As shown in table 14, it is defeated by safe-level SMOTE which has 52 cases. This result may occur because safe-level SMOTE uses the safe-level value to effectively control the suitable location of synthetic instance and ANS1 does not use the entire minority instances to build the model.

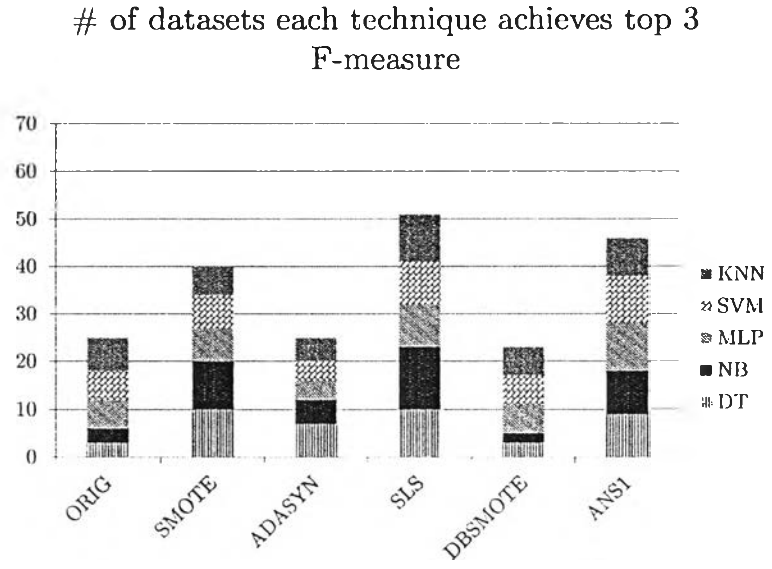


Figure 30: The bar chart of the number of datasets which ANS1 and each oversampling technique achieves the top three F-measure

Table 13: The list of dataset names which ANS1 achieves the best, second best and third best F-measure in each classifier.

Classifier	Datasets in 1st	Datasets in 2nd	Datasets in 3rd	The total number
Decision tree (C4.5)	ecoli, jm1, kc1, vehicle	cm1, kc1, pc1, satimage, yeast		9
Naïve Bayes classifier	kc1	cm1, ecoli, glass, haberman	kc2, optdigits, segment, yeast	9
Multilayer perceptron	jm1, kc1	haberman, optdigits, pc1	cm1, glass, segment, vehicle, yeast	10
Support vector machine	glass, haberman, kc1, yeast	jm1, kc2, pc1	letter, optdigits, satimage	10
K-nearest neighbor	glass, haberman, jm1, kc1, optdigits, vehicle, yeast		cm1	8

Table 14: The number of cases each oversampling technique achieves the F-measure in the ranking 1^{st} -3^{rd}

# of cases as	ORIG	SMOTE	ADASYN	SLS	DBSMOTE	ANS1
1^{st}	17	7	12	7	9	18
2^{nd}	4	20	6	17	8	15
3^{rd}	2	10	9	28	8	13
Total in 1^{st} -3^{rd}	23	37	27	52	25	46

To improve the performance of ANS with the minority outcast handling process, it is expected that ANS with minority outcast handling process or ANS2 provides the better accuracy performance over ANS1. The results are collected in order to compare the average F-measure values from each dataset and classifier to ones from other oversampling techniques. Similar with the previous comparison, the number of cases that each technique achieves the best F-measure and the top three F-measure is ranked and counted. The outcome of ranking is represented as the bar chart counting the number of datasets in figure 31 and figure 32.

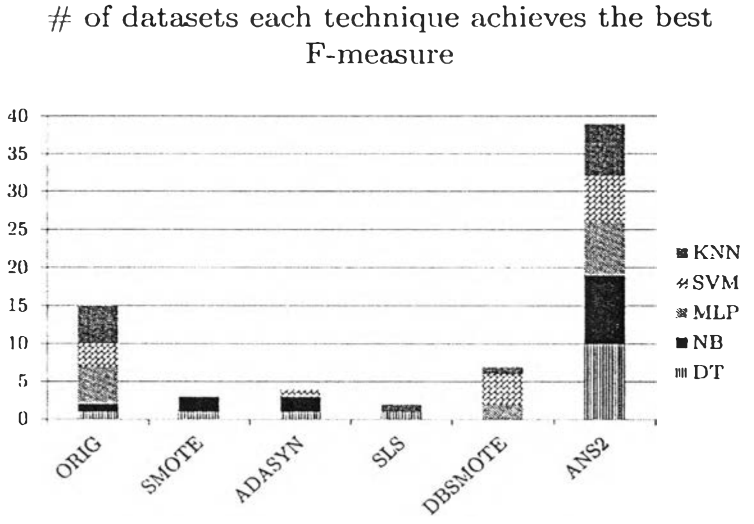


Figure 31: The bar chart of the number of datasets which ANS2 and each oversampling technique achieves the best F-measure

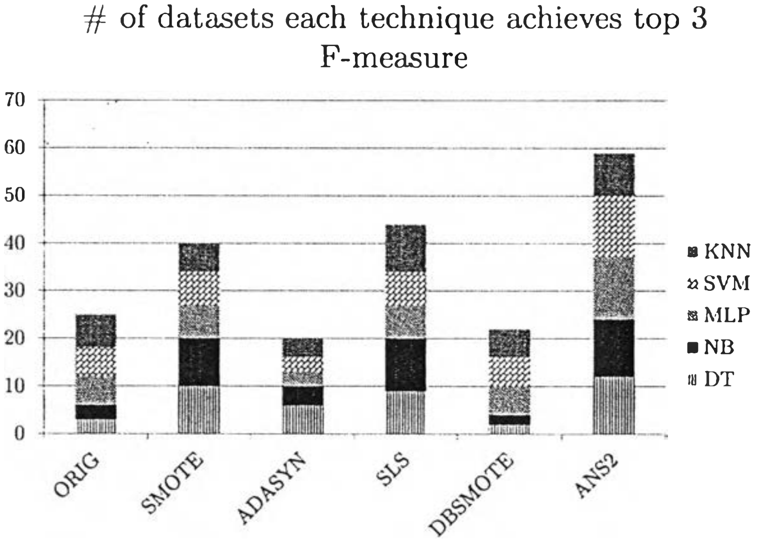


Figure 32: The bar chart of the number of datasets which ANS2 and each oversampling technique achieves the top three F-measure

The bar chart from figure 31 shows the number of datasets which ANS and other oversampling techniques achieve the best F-measure. It shows that ANS2 is much more effective and has the highest number of datasets achieving the best F-measure in every classifier. ANS2 has the best F-measure from 39 cases, the highest and more than the combined number of cases from other oversampling techniques. Moreover, if the scope of consideration is expended to the top three F-measure, figure 32 shows that ANS2 achieves the top three F-measures from over 80 % of cases (57 out of 70) which is the highest percentage among oversampling techniques. The list of dataset names that ANS2 gets the best, second best and third best F-measures in each classifier is shown in table 15 while the number of cases each technique achieves the F-measure in the 1st, 2nd and 3rd rank is shown in table 16.

Table 15: The list of dataset names which ANS2 achieves the best, second best and third best F-measure in each classifier.

Classifier	Datasets in 1st	Datasets in 2nd	Datasets in 3rd	The total number
Decision tree (C4.5)	ecoli, glass, haberman, jm1, kc1, optdigits, pc1, satimage, vehicle, yeast	cm1, letter		12
Naïve Bayes classifier	ecoli, glass, haberman, jm1, kc1, kc2, optdigits, satimage, yeast	cm1	segment, vehicle	12
Multilayer perceptron	glass, haberman, jm1, kc1, letter, optdigits, yeast.	ecoli	cm1, pc1, satimage, segment, vehicle	13
Support vector machine	glass, haberman, jm1, kc1, optdigits, yeast	kc2, pc1, satimage	letter, vehicle	11
K-nearest neighbor	glass, haberman, jm1, kc1, optdigits, vehicle, yeast	ecoli	cm1	9

Table 16: The number of cases each oversampling technique achieves the F-measure in the ranking 1^{st} -3^{rd}

# of cases as	ORIG	SMOTE	ADASYN	SLS	DBSMOTE	ANS2
1^{st}	15	3	4	2	7	39
2^{nd}	4	15	13	21	10	7
3^{rd}	4	19	6	23	7	11
Total in 1^{st} -3^{rd}	23	37	23	46	24	57

Similar with RSLS, the Wilcoxon signed-rank test with ANS1 and ANS2 is also conducted in order to verify whether the positive differences of F-measure caused by ANS1 and ANS2 against other oversampling techniques are significant. In the first part of this section, F-measure values from each round of experiments from ANS1 and ANS2 are compared with F-measure value from other oversampling techniques in the same sampling training and test set and the same classifier. Then, the positive median of difference in each comparison is tested for its significance. The null hypothesis of each test is set as the median of difference is less than or equal to zero. The confidence level is set at 95%, so if a p-value of test is less than 0.05, then the alternative hypothesis which is the median of difference between the controlled algorithm (ANS1 and ANS2, respectively) and the other compared algorithm are positive. Using the results from every round of experiments regardless of various classifiers and datasets, the test result is shown in the table below.

Table 17: The Wilcoxon signed-rank of the difference of F-measure from ANS1 and ANS2 against other oversampling techniques

ANS1 against	Median of Difference	p-value	ANS2 against	Median of Difference	p-value
Original	0.0321	7.1435×10^{-116}	Original	0.0459	4.9496×10^{-156}
SMOTE	0.0051	1.0386×10^{-09}	SMOTE	0.0176	6.4233×10^{-75}
ADASYN	0.0138	1.2541×10^{-67}	ADASYN	0.0271	1.4807×10^{-151}
SLS	0.0003	7.3270×10^{-01}	SLS	0.0109	2.3378×10^{-41}
DBSMOTE	0.0155	9.1304×10^{-71}	DBSMOTE	0.0292	2.4914×10^{-139}
			ANS1	0.0170	6.5542×10^{-124}

It can be seen from the results in table 17 that ANS1 and ANS2 both achieve significantly positive differences against other oversampling techniques except when ANS1 is compared against SLS which is the only case that its p-value is more than 0.05. As already seen in table 14, safe-level SMOTE has slightly more cases achieving the top three than ANS1, so this result is expected. But after adding the minority outcast handling process as in ANS2, it could be seen that ANS2 has significantly positive difference of F-measure against all other oversampling techniques as p-values are all less than 0.05.

Table 18: The Wilcoxon signed-rank of the difference of F-measure from ANS1 and ANS2 against other sampling techniques in each classifier

Classifier	ANS1 against	Median of Difference	p-value	ANS2 against	Median of Difference	p-value
DT	ORIG	0.0241	2.1900×10^{-26}	ORIG	0.0359	3.4400×10^{-42}
	SMOTE	0.0011	0.1718	SMOTE	0.0137	3.2000×10^{-17}
	ADASYN	0.0011	0.1209	ADASYN	0.0137	1.4600×10^{-17}
	SLS	0.0011	4.1461×10^{-02}	SLS	0.0132	5.4700×10^{-18}
	DBSMOTE	0.0161	2.1800×10^{-21}	DBSMOTE	0.0299	1.3900×10^{-45}
				ANS1	0.0251	3.2400×10^{-61}
NB	ORIG	0.0208	2.5800×10^{-27}	ORIG	0.0437	4.8000×10^{-49}
	SMOTE	-0.0098	1	SMOTE	0.0089	9.8767×10^{-02}
	ADASYN	0.0108	3.3100×10^{-10}	ADASYN	0.0286	8.8800×10^{-28}
	SLS	-0.0041	1	SLS	0.0122	3.5300×10^{-07}
	DBSMOTE	0.0603	5.0300×10^{-80}	DBSMOTE	0.0876	5.6600×10^{-97}
				ANS1	0.0263	6.0900×10^{-75}
MLP	ORIG	0.0258	1.3200×10^{-09}	ORIG	0.0308	1.7600×10^{-15}
	SMOTE	0.0076	1.0700×10^{-06}	SMOTE	0.0230	1.4400×10^{-24}
	ADASYN	0.0188	1.7100×10^{-22}	ADASYN	0.0306	2.2600×10^{-46}
	SLS	0.0001	0.26821	SLS	0.0110	5.5500×10^{-13}
	DBSMOTE	0.0095	1.0540×10^{-04}	DBSMOTE	0.0174	1.1900×10^{-15}
				ANS1	0.0122	7.1400×10^{-13}
SVM	ORIG	0.1042	5.2200×10^{-65}	ORIG	0.1066	3.8800×10^{-67}
	SMOTE	0.0158	1.1100×10^{-27}	SMOTE	0.0274	4.6100×10^{-52}
	ADASYN	0.0244	7.0300×10^{-35}	ADASYN	0.0372	3.6600×10^{-58}
	SLS	0.0042	2.0819×10^{-03}	SLS	0.0117	3.1700×10^{-20}
	DBSMOTE	-0.0020	0.67452	DBSMOTE	0.0063	9.6200×10^{-06}
				ANS1	0.0179	3.7900×10^{-56}

Classifier	ANS1 against	Median of Difference	p-value	ANS2 against	Median of Difference	p-value
KNN	ORIG	0.0143	3.9300×10^{-10}	ORIG	0.0192	2.4600×10^{-09}
	SMOTE	0.0094	7.5700×10^{-13}	SMOTE	0.0162	1.2500×10^{-12}
	ADASYN	0.0164	2.8700×10^{-25}	ADASYN	0.0220	4.1800×10^{-21}
	SLS	0.0014	0.10058	SLS	-0.0004	2.8819×10^{-02}
	DBSMOTE	0.0032	2.7500×10^{-05}	DBSMOTE	0.0148	9.1800×10^{-07}
				ANS1	-0.0020	0.42596

Based on these Wilcoxon signed-rank test results separated by classifiers shown in table 18, ANS1 cannot achieve positive difference against some oversampling techniques in some classifier. For decision tree, ANS1 has p-values larger than 0.05 when it is compared against SMOTE and ADASYN and barely smaller than 0.05 against safe-level SMOTE. This could imply that ANS1 cannot overcome SMOTE, ADASYN and safe-level SMOTE clearly in this classifier. However, the test with ANS2, the p-values against every oversampling technique is less than 0.05. So, this can be concluded that ANS2 has better performances than other oversampling techniques in this classifier. In naïve Bayes classifier, ANS1 has p-values larger than 0.05 and negative difference against SMOTE and safe-level SMOTE. This shows that it cannot provide better performance over these two oversampling techniques. However, ANS2 still provides a better performance against all oversampling techniques in this classifier. This could mean minority outcast handling help improving the classification performance of ANS1 in naïve Bayes classifier.

For multilayer perceptron, the similar situation occurs as ANS1 can provide the positive difference in almost every comparison with other oversampling technique significantly except safe-level SMOTE. ANS2 which minority outcast handling process is included can overcome and has positive difference in every test against other oversampling techniques with p-values less than 0.05. ANS1 also provides the positive difference in every test against other oversampling techniques except DBSMOTE when it is trained with support vector machine. It requires the minority outcast handling to achieve the positive difference against every oversampling technique significantly. However, ANS2 does not work better than safe-level SMOTE significantly whether minority outcast handling is included in *k*-nearest neighbor, as it got p-value more than 0.05 against safe-level SMOTE while it can achieve p-value less than 0.05 against any other oversampling techniques.

The concern in this dissertation is whether adaptive neighbor process or minority outcast handling process is a factor for the improving performance. To answer this concern, experiments on 70 cases of 14 UCI datasets and 5 classifiers from SMOTE whose outcasts are removed (SMOTEO) and adaptive neighbor SMOTE (ANS) are performed. There are two versions of SMOTEO and ANS in this experimental setting, i.e., ones without minority outcast handling (SMOTEO-1 and ANS1) and one with minority outcast handling (SMOTEO-2 and ANS2). The results represented as the average F-measure values are reported in table 27 and summarized in table 19. The parameter *k* is set as 5 which is the setting used for SMOTE in the original paper of SMOTE and related research papers. The results from techniques that do not apply minority outcast handling, i.e., SMOTEO-1 and ANS1, are paired. The number of cases which ANS1 achieves higher F-measure than one from SMOTEO-1 is 37 which is more than half of total cases. Similarly, the results from two techniques that apply minority outcast handling, i.e., SMOTEO-2 and ANS2, are also paired. The number of cases which ANS2 achieves higher F-measure than one from SMOTEO-1 is 40. The number of cases which SMOTEO-1 has better F-measure than ANS1 but ANS2 has better F-measure than SMOTEO-2 is only 6. This implies that there are only few cases which minority outcast can overturn the result between these two oversampling techniques. Most cases (34) that ANS can overcome SMOTE happen when ANS1 has already higher F-measure than SMOTEO-1. So, it can conclude that adaptive neighbor SMOTE can provide the better classification performance over original SMOTE with its dynamic *k* process. Moreover, when ANS is more effective than SMOTE, minority outcast process helps improving the result further.

Table 19: The number of cases which averaged F-measure of ANS1 or ANS2 is higher/lower than one of SMOTEO-1 or SMOTEO-2.

	ANS1 > SMOTEO-1	SMOTEO-1 > ANS1	Total by rows
ANS2 > SMOTEO-2	34	6	40
SMOTEO-2 > ANS2	3	27	30
Total by columns	37	33	70

To further investigate the effect of adaptive neighbor approach and minority outcast handling, the analysis of variance is also performed on this experimental result. The significant level for this test is set as 0.95 which means that if the p-value is less than 0.05, there is significant difference between the F-measure mean of two groups. The ANOVA result between the F-measure values from SMOTE with the fixed *k* = 5 and the ones from ANS is shown in table 20.

Table 20: The ANOVA table between F-measure values from SMOTE with the fixed k = 5 and the ones from ANS

	Df	Sum Square	Mean Square	F value	Pr(> F)
SMOTEO vs ANS	1	0	0.00053	0.011	0.916
Residuals	13998	663	0.04736		

The ANOVA table shown in table 20 displays that the p-value is 0.916 which is more than the critical value 0.05. It implies that the mean of F-measure from SMOTE with the fixed *k* = 5 and adaptive neighbor SMOTE is not significantly different. However, the fixed *k* = 5 is required the tuning of a parameter k in order to find the optimal value which costs more time and resources than adaptive neighbor while yielding the similar overall classification result based on ANOVA.

The effect of minority outcast handling applied in SMOTE and adaptive neighbor SMOTE is also investigated by ANOVA. With the significant level at 0.05, the ANOVA test is conducted to compare the mean of F-measure from two groups, ie, a group of sampling techniques without applying the minority outcast handling and a group of sampling techniques applying the minority outcast handling. The result of ANOVA is shown in table 21.

Table 21: The ANOVA table between F-measure values from oversampling techniques without applying minority outcast handling and the ones with minority outcast handling

	Df	Sum Square	Mean Square	F value	Pr(> F)
w/o vs with outcast	1	0.6	0.5776	12.21	4.78×10^{-4}
Residuals	13998	662.4	0.0473		

The ANOVA table shown in table 21 displays that the p-value is 4.78×10^{-4} which is less than 0.05. This implies that the mean of F-measure from oversampling techniques without minority outcast handling and oversampling techniques with minority outcast handling is significantly different. This can be affirmed that minority outcast handling can effectively enhance the classification performance in the class imbalance problem with the significant improvement.