19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

Efficiency Comparisons Between K-Centers and K-Means Algorithms

Varin Chouvatut, Wattana Jindaluang*, Ekkarat Boonchieng, Thapanapong Rukkanchanunt
The Theoretical and Empirical Research Group
Center of Excellence in Community Health Informatics
Department of Computer Science
Faculty of Science, Chiang Mai University
Chiang Mai, Thailand
varinchouv@gmail.com, wjindaluang@gmail.com (*corresponding author), ekkarat@boonchieng.net

Abstract—**This paper proposes an under-sampling method with an algorithm which guarantees the sampling quality called k-centers algorithm. Then, the efficiency of the sampling using under-sampling method with k-means algorithm is compared with the proposed method. For the comparison purpose, four datasets obtained from UCI database were selected and the RIPPER classifier was used. From the experimental results, our under-sampling method with k-centers algorithm provided the Accuracy, Recall, and F-measure values higher than that obtained from the under-sampling with k-means algorithm in every dataset we used. The Precision value from our k-centers algorithm might be lower in some datasets, however, its average value computed out of all datasets is still higher than using the under-sampling method with k-means algorithm. Moreover, the experimental results showed that our under-sampling method with k-centers algorithm also decreases the Accuracy value obtained from the original data less than that using the under-sampling with k-means algorithm.**

Keywords—k-centers algorithm; imbalance dataset; under-sampling

I. INTRODUCTION

Imbalance dataset is a dataset whose numbers of data items retrieved from different groups are not equal (imbalanced). Mostly, the number of data items of our interest is extremely less than the number of other data items. The main goal of classification problem is to classify such kind of mentioned data items. Thus, the group or class of the small number of data items of our interest is called minority class and the rest which has a large number would be majority class. Examples of application and research relating to data with characteristics of this imbalance dataset are as explained in [1-5]. Researchers in [1] compared performance of various methods used to handle with imbalance dataset in micro-calcification classification. The classification is very important in breast-cancer examination. In [2], Efficiency of over-sampling, SMOTE, and under-sampling techniques were examined in order to balance the cardiovascular data. In addition, the research proposed using an under-sampling technique to the data. Their experimental results showed that the proposed technique provided better efficiency than other traditional technique. The research of [3] proposed a technique to handle with an imbalance dataset using the compact evolutionary interval-valued fuzzy rule-based method. The method does not require any data preprocessing or any data sampling. Data used in the research is for financial applications. Their experiments demonstrated that their proposed technique gave better efficiency than the C4.5 decision tree, type-1, or interval-valued fuzzy counterparts which use SMOTE in data preprocessing. Researchers in [4] handled the imbalance dataset by storing objects in Euclidean n-space and then classifying the new incoming data using the distance calculation between the new data and their nearest generalized exemplar. Their experiments showed that the researchers' approach is better than the contemporary methods. In [5], researchers compared the Support Vector Machine (SVM), Decision Tree, and Naïve Bayes algorithms with data from UCI database [6]. The comparison was done in terms of sensitivity, specificity, G-means, and time-based efficiency. Their research found that SVM gave higher efficiency than the other two algorithms with respect to the sensitivity or the specificity in all datasets. And the SVM gave higher G-means when it was applied to large datasets. The problems usually found when one uses a classifier for a balance dataset to classify data items of an imbalance dataset is that the classifier traditionally used for a balance dataset will know only data items mostly of the majority class; while the small number of data items of our interest from the minority class will still be unknown and, thus, be considered as outliers. From the problem of ignoring the small number of data items in minority class, the traditional classifier will predict the group or class of a new coming data item a majority class though the new data item is exactly of a minority class. There are two approaches which can handle this imbalance dataset. One approach is to add data items to the minority class to gain likely equal number of data items to the majority class, this approach is called over-sampling method. Inversely, another approach is to reduce the number of data items of the majority class to get likely equal number of data items to the minority class, this approach is called under-sampling method. Both approaches have different pros and cons. In this paper, we chose the under-sampling method and we applied the cluster-based algorithm to the process of selecting the representative data items out of the majority class.

A popular cluster-based algorithm is k-means as used in [7-10]. The algorithm works with iterative processes. For each

978-1-4673-7825-3/15/$31.00 ©2015 IEEE

iteration, it will try to search for a centroid of data cluster which minimizes the mean distance between the centroid to the data items within its cluster. If the current iteration provided a better centroid, the newly better centroid will be substituted for the cluster's centroid. The iterative algorithm will stop when centroid of each cluster has no significant change. One main problem usually found using k-means algorithm is because the algorithm works iteratively without a definite stop and the iterative condition principally depends on the initial positions of the centroid and characteristics of the data. An instance of this first problem caused by the initial centroid positions and the data characteristics can be a case of having scattered data viewing in two dimensions; if the initial positions of centroids were randomly chosen and the random positions were clustered around each other, in this case, the k-means algorithm will spend a huge number of iterations. One more problem is that the finally obtained centroid may be only a pseudo-centroid which is not a real data in the dataset but the synthetic data. Thus, if one wants to apply the obtained centroids to an application in order to analyze other input data, he may face a serious problem due to the not-real data using.

Consequently, we chose to use another cluster-based algorithm called k-centers algorithm to select the representative data items from the majority class for our under-sampling method. Aim of the k-centers algorithm is to choose centers of the data clusters each of which minimizes the longest distance amongst the data items in a cluster and its center. K-centers algorithm has many advantages. The main advantage of k-centers algorithm is that it has certain processes, not like an iterative algorithm as k-means, in other words, k-centers algorithm has a definite stop. And the center obtained from the k-centers algorithm is exactly the real data in the dataset. Moreover, the k-centers algorithm is in the group of algorithms with the support of theorem guaranteeing that quality of the obtained centers is very close to the optimum solution.

Since we used k-centers algorithm in sampling data out of the majority class to get the number of data items almost equal to the minority class, we thus measured the efficiency of our under-sampling method with four datasets from a standard database named UCI [6] and used RIPPER as a classifier for the purpose of efficiency comparison. From the classification process, the Precision, Recall, F-measure, and Accuracy values were then computed.

This paper is divided into sections as follows. Section II explains the main concept of under-sampling method using k-centers algorithm. Section III shows experimental results and their discussion and conclusion are explained in section IV.

II. Under-sampling with K-centers Algorithms

The under-sampling method has an obvious advantage over the over-sampling method which is its significantly faster processing time. Unfortunately, its important disadvantage is that if the approach of selecting the representative data items was not good enough, data items with some important information may not be selected from the dataset and thus the information will be lost and ignored in further classification. For example, if we chose randomization to sample data items out of the majority class to be the representative data of the class, the random data items may be obtained from some clustered data items which are not scattered enough around the class. In such case, the sampled data cannot be counted as a good representative data group of the majority class.

K-centers algorithm is a method to sample the representative data out of a dataset. The representative data items selected are called centers of clusters where there will be k items using k-centers algorithm. The k-centers algorithm were first proposed in the research of [11]. In the research, other than proposing the selective algorithm, it also prove a theorem which guaranteed the quality of the result in that the representative data selected by k-centers algorithm would provide the cost of not more than double of the cost from optimum solution. This means if we measured the similarity of two sets of data (the original data and their representative samples) using the distance, k-centers algorithm will still retain the important information of the original dataset in minority class.

We thus used k-centers algorithm to sample data items from the majority class by determining k which is the number of data items to be selected from the majority class in each iteration to $\beta\gamma$, where β is the number of data items in minority class and initially set γ to $\alpha - 1$ when α is the imbalance ratio which is the ratio of the number of data items in majority class to the number of data items in the minority class. After that, all selected data items would be combined with the data items of the minority class. Next, process a five-fold cross validation and compute the Precision, Recall, F-measure, and Accuracy values. Then, reduce γ by 1 and repeat the k-centers algorithm again. The under-sampling method will process its last iteration when $\gamma < 2$ (closely to 1), that is, the number of data items of the majority class is assumed to be almost equal to the minority class.

III. Expermental Results

Experiments in this paper worked with four datasets from UCI [6] composed of Glass Identification, Statlog, Page-Blocks, and E-coli. The characteristics of the datasets are concluded in Table I. Note that RIPPER was chosen as the classifier and implemented using JRIP in WEGA.

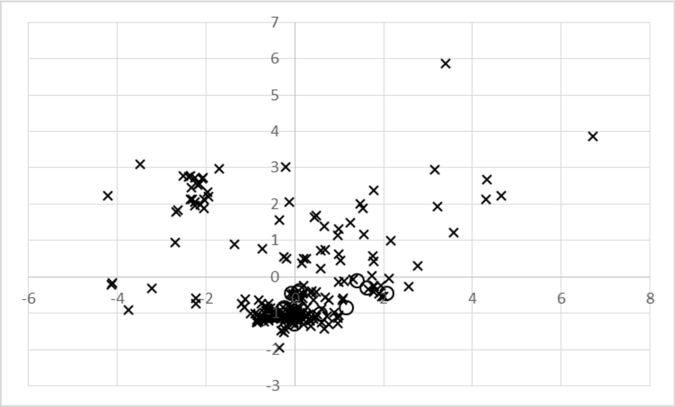
TABLE I. CHARACTERISTICS OF FOUR DATASETS

Dataset Name	No. of Instances	No. of Attributes	Minority Class Name	Imbalance Ratio (α)
Glass Identification	214	10	vehicle windows	11.58
Statlog	6,435	36	damp grey soil	9.69
Page-Blocks	5,743	10	picture	46.59
E-coli	336	8	imU	8.60

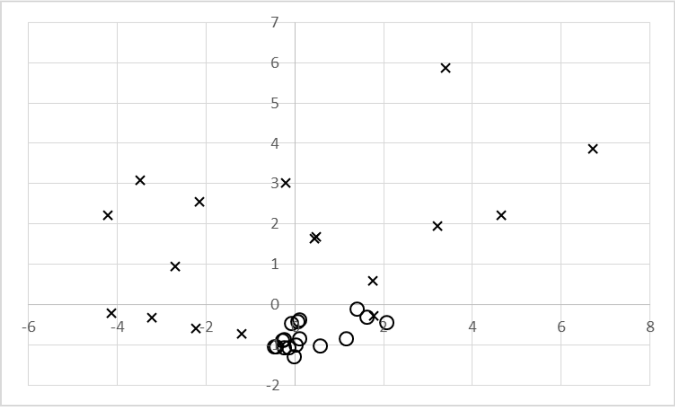
Although our sampling method used various k's, from our experimental results, the k appropriate to generate a training set where the Precision, Recall, and F-measure values increased while the Accuracy value decreased acceptably is $k \approx \beta$. In other words, $\gamma \approx 1$ which means sampling data items out of the majority class with an almost equal number of the minority class. So comparing the efficiency of under-sampling between using k-centers algorithm and using k-means algorithm was done when k's of the two algorithms are close to β only.

To compare the selection of the representative data using k-centers and k-means algorithms, we reduce the number of data attributes in each dataset to two using Principal Component Analysis (PCA). Then, the representative data selected by the two algorithms were plotted in a 2D graph in Figs. 1 – 4. Note that, the representative data shown in the 2D graphs is the attribute-reduced data of the original and real data whose attributes (dimensions) were reduced to only two to be shown in the graphs, thus, the illustration of the reduced-dimension data may cause some distortions to position appearance in the graphs.

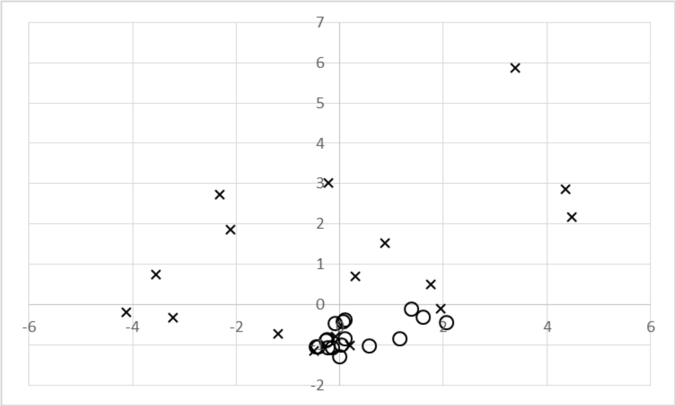
In Figs. 1 – 4, data items of the majority class sampled by an under-sampling method, using either k-centers or k-means algorithm, are represented with the cross marks (×) and the whole data items of the minority class are represented with the round marks (○). Note that the minority class are the same in all figures.



(a)



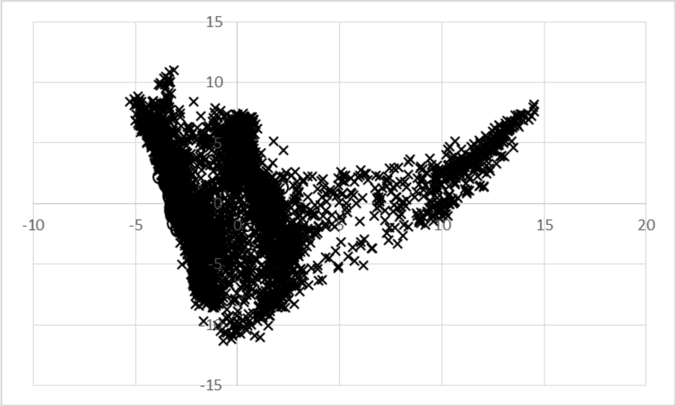
(b)



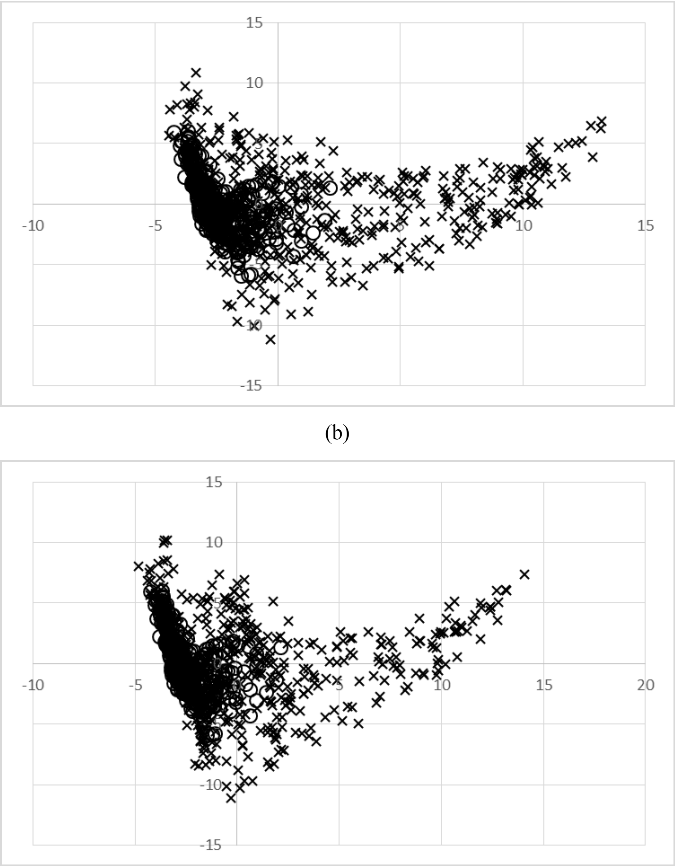
(c)

Fig. 1. Glass-Identification dataset; (a) original, (b) under-sampling with k-centers, and (c) under-sampling with k-means.

From Figs 1 – 4, data sampled by k-centers algorithm will be all real data while, for sampling data by k-means algorithm, most sampled data will be pseudo-centroids. Such as in Glass-Identification dataset, the pseudo-centroids of the dataset are (0.2202, -1.0064), (4.48424, 2.17556), (-3.553772, 0.75805) etc. For Statlog dataset, the pseudo-centroids are as (-4.82059, 8.035302), (0.11384, 2.959885), (14.07302, 7.383806), etc. In Page-blocks dataset, the pseudo-centroids include (-0.91194, -0.43188), (0.61092, 3.55454), (-0.08852, -2.73608), etc. And for E-coli dataset, the pseudo-centroids are as (0.415794, -0.63254), (0.400749, 0.264159), (0.647237, -0.65661), etc.



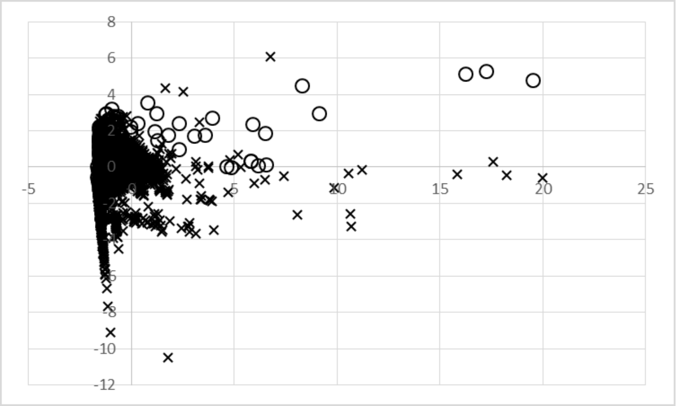
(a)



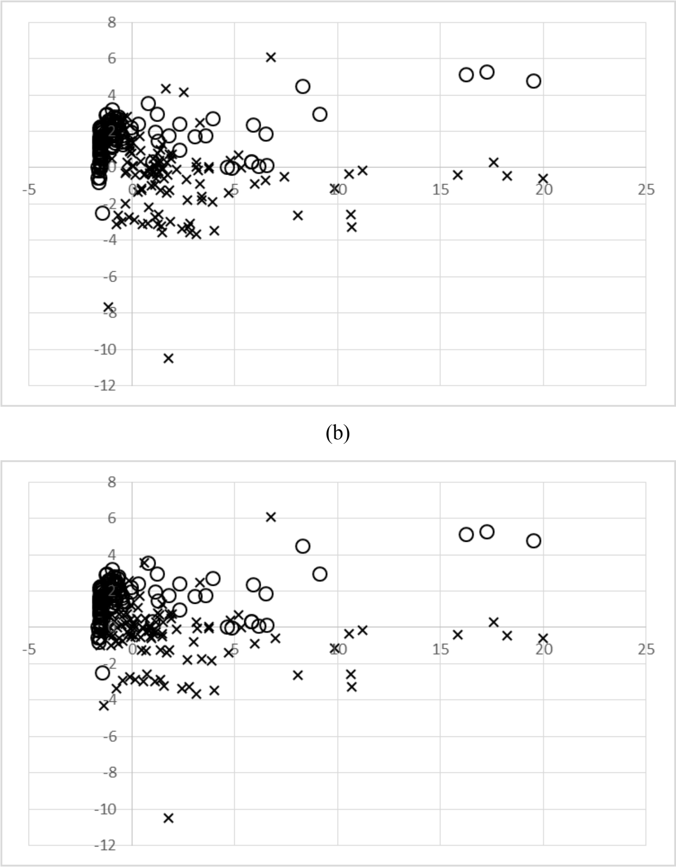
(b)

(c)

Fig. 2. Statlog dataset; (a) original, (b) under-sampling with k-centers, and (c) under-sampling with k-means.



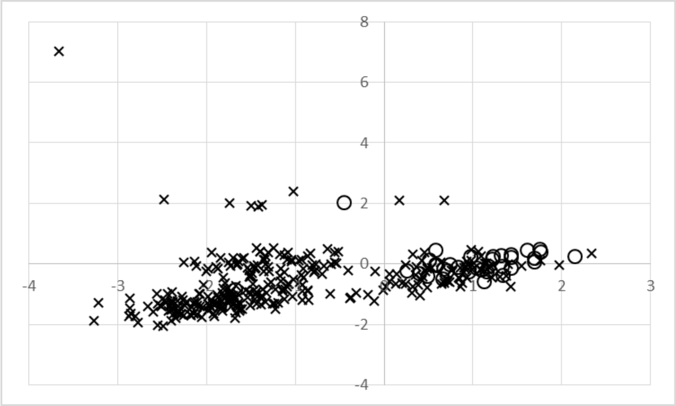
(a)



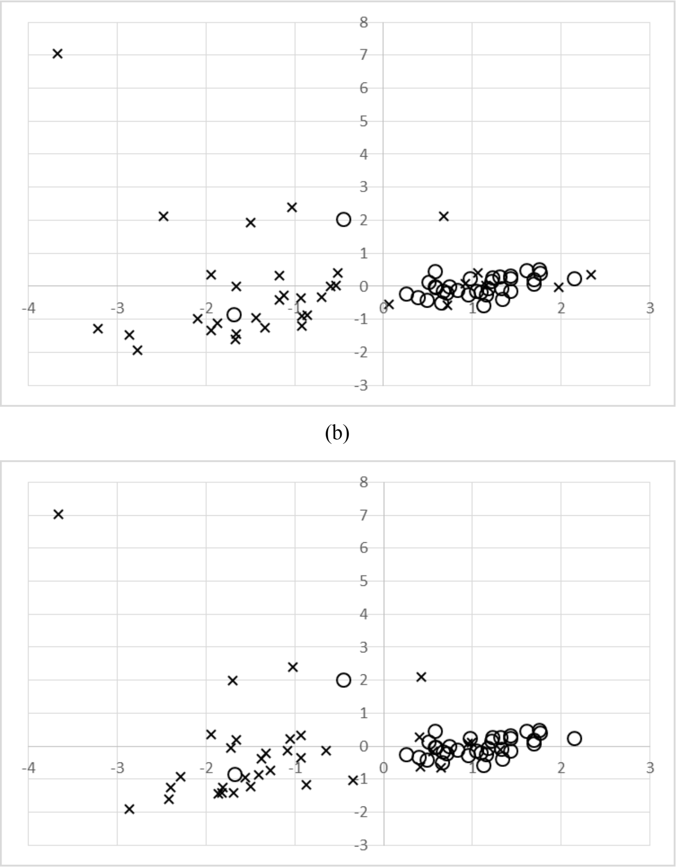
(b)

(c)

Fig. 3. Page-blocks dataset; (a) original, (b) under-sampling with k-centers, and (c) under-sampling with k-means.



(a)



(b)

(c)

Fig. 4. E-coli dataset; (a) original, (b) under-sampling with k-centers, and (c) under-sampling with k-means.

The Accuracy values of the four datasets computed from the under-sampling methods using k-centers and k-means algorithms were compared in Table II.

TABLE II. ACCURACY OF FOUR DATASETS

Dataset Name	Accuracy	
	k-centers	*k-means*
Glass Identification	91.1765	85.2941
Statlog	88.1928	83.4940
Page-Blocks	90.4348	90.0000
E-coli	80.0000	80.0000

From Table II, the under-sampling method using k-centers algorithm gave the Accuracy value not smaller than using k-means algorithm in all four datasets where k-centers gave the average Accuracy of 87.4510% and k-means gave 84.6970% in average.

Table III shows the comparisons of decreases in Accuracy from the original datasets due to the two under-sampling methods.

TABLE III. PERCENTAGES OF ACCURACY DECREASING FROM THE ORIGINAL DATASET

Dataset Name	Percentage of Accuracy Decreasing from the Original Dataset	
	k-centers	*k-means*
Glass Identification	0.9555	7.3455
Statlog	4.4848	9.5737
Page-Blocks	8.3257	8.7664
E-coli	11.8689	11.8689

From Table III, the under-sampling method using k-centers algorithm caused the Accuracy value of the representative data items to decrease from the original dataset less than the under-sampling method using k-means algorithm for all datasets except the E-coli dataset where two algorithms provided the same Accuracy. K-centers and k-means algorithms gave the average Accuracy of 6.4087% and 9.3886%, respectively.

The Precision, Recall, and F-measure values from the under-sampling methods using both k-centers and k-means algorithms are shown in Tables IV – VI, respectively.

TABLE IV. PRECISION OF FOUR DATASETS

Dataset Name	Precision	
	k-centers	*k-means*
Glass Identification	0.8889	0.8333
Statlog	0.8801	0.8294
Page-Blocks	0.8974	0.9259
E-coli	0.7692	0.7692

From Table IV, the Precision values from the under-sampling method using k-centers algorithm was higher than from k-means algorithm in two datasets, Glass Identification and Statlog but lower than k-means in Page-Blocks dataset while equal to k-means in E-coli dataset. However, the Precision of k-centers is still higher than k-means in average, that is, k-centers gave 0.8589 and k-means gave 0.8394.

TABLE V. RECALL OF FOUR DATASETS

Dataset Name	Recall	
	k-centers	*k-means*
Glass Identification	0.9412	0.8824
Statlog	0.8843	0.8434
Page-Blocks	0.9130	0.8696
E-coli	0.8571	0.8571

From Table V, the under-sampling method using k-centers algorithm gave Recall values not less than the k-means algorithm in every dataset where k-centers and k-means algorithms gave 0.8989 and 0.8631, respectively, in average.

TABLE VI. F-MEASURE OF FOUR DATASETS

Dataset Name	F-measure	
	k-centers	*k-means*
Glass Identification	0.9143	0.8571
Statlog	0.8822	0.8363
Page-Blocks	0.9052	0.8969
E-coli	0.8108	0.8108

From Table VI, the under-sampling method using k-centers algorithm gave F-measure values not less than the k-means algorithm in all datasets where the average F-measure values are 0.8781 and 0.8503 for k-centers and k-means, respectively.

IV. Discussion and Conclusion

The under-sampling method is one of the approaches which can be used to handle with the imbalance datasets by reducing the number of data items in majority class to be equal to the size of minority class. Although the under-sampling method has an advantage of shorter time in training comparing with the over-sampling method, the usually found problem of the under-sampling is that if the selection of representative data items out of the majority class is not enough efficient, some important information may be lost and consequently the classification process will gain less efficiency also.

To avoid the problem, we thus proposed the algorithm with the support of theorem guaranteeing the quality of representative-data selection, in that, with the guaranteed algorithm, if the similarity of the representative data and the original dataset was represented by the distance between data items, thus the cluster-based approximation algorithm named k-centers can reduce the drawback caused by the problem.

Thus, we handled the imbalance dataset by under-sampling the dataset using k-centers algorithm to select the representative data out of the majority class with various imbalance ratios. Then, combined the selected data items of the majority class with the minority class to generate a training set. The sampling process would stop when the number of representative data is almost equal to the number of data in minority class.

We experimented with four datasets retrieved from a standard database named UCI and used RIPPER as the classifier. Efficiency of the under-sampling method using k-centers algorithm was then compared with the under-sampling method using a popular algorithm named k-means. From the efficiency comparisons, the Accuracy, Recall, and F-measure values obtained from k-centers algorithm are not less than from k-means in all datasets, even more, their average values are higher than those of the k-means. Some Precision values of the k-centers are a little lower than that of the k-means in some datasets, however, the average Precision value of k-centers from all datasets is still higher than that of k-means. Moreover, we found that the method using k-centers algorithm also reduced the Accuracy value of the representative data selected from the original dataset less than the Accuracy value obtained from using the k-means algorithm.

References

[1] M. Kumar and H.S. Sheshadri, “On the classification of imbalanced datasets”, International Journal of Computer Applications 44 (2012):1-7.

[2] M.M. Rahman and D.N. Davis, “Addressing the class imbalance problem in medical datasets”, International Journal of Machine Learning and Computing 3 (2013):224 - 228.

[3] J.A. Sanz, E. Bernardo, F. Herrera, H. Bustince and H. Hagras, “A compact evolutionary interval-valued Fuzzy Rule-Based classification system for the modeling and prediction of real-world financial applications with imbalanced data”, IEEE Transactions on Fuzzy Systems, in press.

[4] S. Garcia, J. Derrac, I. Trguero, C.J. Carmona and F. Herrera, “Evolutionary-based selecion of generalized instances for imbalanced classification”, Knowledge-Based Systems 25 (2012): 3 – 12.

[5] S. Zhang, S. Sadaoui and M. Mouhoub, “An empirical analysis of imbalanced data classification”, Computer and Information Science 8 (2015):151 – 162.

[6] UCI Database, https://archive.ics.uci.edu/ml/datasets.html (searched in Jul. 2014).

[7] M.H. Zafar and M. Ilyas, “A clustering based study of classification algorithms”, International Journal of Database Theory and Application 8 (2015): 11 – 22.

[8] P. Nagchoudhury and K. Choudhary, “Classification of swarm intellignece based clustering methods”, International Journal of Computer Applications 91 (2014): 28 – 33.

[9] W. Prachuabsupaki and N. Soonthornphisa, “Clustering and combined sampling approaches for multi-class imbalanced data classification”, Advances in Information Technology and Industry Applications Lecture Notes in Electrical Engineering, 136(2012), pp 717-724.

[10] Y. Zhang, L. Zhang and Y. Wang, “Cluster-based majority under-sampling approached for class imbalanced learning”, Information and Financial Engineering (ICIFE) 2 nd IEEE International Conference on, 17-19 Sept. 2010, pp. 400-404.

[11] D.S. Hochbaum and D.B. Shmoys, “A unified approach to approximation algorithms for bottleneck problems”, Journal of the ACM, 33, pp 533-550, 1986.