2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), and 2015 IEEE 12th International Conf on Embedded Software and Systems (ICESS)

Multi-Classes Imbalanced Dataset Classification Based on Sample Information

Chuang Yu
School of Software Technology
Dalian University of Technology
Dalian, China
yuchuang@mail.dlut.edu.cn

Fengqi Li
School of Software Technology
Dalian University of Technology
Dalian, China
lifengqi@dlut.edu.cn

Guangming Li
School of Software Technology
Dalian University of Technology
Dalian, China
guangmingli@mail.dlut.edu.cn

Nanhai Yang
School of Software Technology
Dalian University of Technology
Dalian, China
nanhai@dlut.edu.cn

***Abstract*—The classification boundary for multi-classes imbalanced dataset is difficult to judge, posing an important challenge on classification methods. Aiming at this problem, we propose an multi-classes imbalanced data classification algorithm based on sample information. The proposed algorithm applies the sample information measurement to multi-classes imbalanced dataset. Furthermore, a classifier is devised to classify the data. Experiments on IRIS, WINE, GLASS datasets show that our proposed scheme produces a promising result for classifying multi-classes imbalanced data.**

Keywords—Sample Information; Multi-classes; Imbalanced Datasets Classification; Resampling

I. INTRODUCTION

In recent years, more and more imbalanced datasets have been produced in various applications. For example, in the network intrusion monitoring, abnormal accesses occur with small probability, while most accesses are normal. Such that, the number of data samples of normal accesses is much larger than that of abnormal accesses. Likewise, in protein or gene abnormality detection applications, the ratio of mutant protein or gene to normal protein or gene is minimal. In this condition, there exist more normal protein or gene samples than mutant ones. Although the number of abnormal samples is small in these applications, how to determine such abnormal samples plays important role in our daily life. Furthermore, this imbalanced situation is not only in binary-classification problems, but also commonly occurred in multi-class datasets.

Imbalanced datasets refers to the datasets with different sample numbers in different classes [3]. The class with more samples is called majority class, while the other one is called minority class, as Fig. 1 shows.

The imbalance ratio characterizes the imbalanced degree of datasets. In the binary-classification situation, we regard the ratio of number of majority class samples to that of minority class as the imbalance ratio, and the larger the imbalance ratio, the more imbalanced degree of datasets. The imbalance ratio in some datasets can reach 100:1, while some others even get 10000:1. However, for multi-class datasets, the imbalanced degree has no uniform standards.

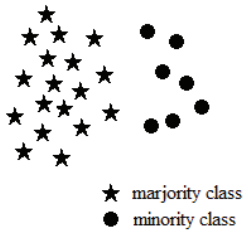

Fig. 1. Illustration of Imbalanced Dataset

Since most classification methods are proposed on the basis of the assumption that different classes contain the almost the same number of data samples, once occurs the imbalanced situation, classification boundaries trained by these classification methods will deviate to the minority class samples, which will leads the classifier classifies some minority class samples to the majority class, and then the classification accuracy will decrease. In Fig. 2, the classification boundary deviates to the minority class.

The task of imbalanced datasets classification can be solved by Resampling [4] (Resampling) method. Resampling methods are used to change the proportion of class distribution to balance the dataset. Resampling methods can be classified in two forms: oversampling or undersampling. Oversampling increase the number of samples in minority class; undersampling decrease the number of samples in majority class in contrary. Each technique has its advantages and

978-1-4799-8937-9/15 $31.00 © 2015 IEEE
DOI 10.1109/HPCC-CSS-ICESS.2015.327

IEEE computer society

Authorized licensed use limited to: UNIVERSITY TENAGA NASIONAL. Downloaded on May 14,2023 at 17:03:47 UTC from IEEE Xplore. Restrictions apply.

disadvantages; oversampling technique adds new synthetic samples into the dataset, leading to over-fitting problem. Undersampling technique discards the sample of majority classes in order to achieve balance, leading to lose information of datasets.

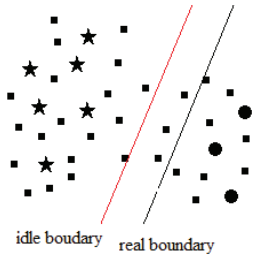

Fig. 2. Classification Boundary Deviates

Sample information to express the degree of sample's contribution to the classification results has an important influence on the resampling technique, it determines in which produce synthetic samples around samples. Most resampling methods consider that the information of samples nearby classification boundary is higher; therefore, to generate artificial samples around these samples has a greater contribution to the classification results. The traditional resampling methods are mostly used in binary-classification datasets, in whereas the studies of imbalanced classification on multi-classes imbalanced datasets are very few [5]. For multi-classes dataset classification, classification boundary is often difficult to determine. Therefore, we propose a sample information measurement method based on local density, and apply it in resampling methods on multi-classes imbalanced datasets.

In addition, when applying resampling methods on imbalance datasets, they need to identify the imbalance ratio of datasets to stop the sampling process. Therefore, we propose an imbalance ratio measurement to describe the imbalanced degree of unbalanced datasets.

II. Relate Work

Oversampling technique has been more widely used by its better use of the known datasets. The simplest sampling method is random oversampling. Its principle is very simple, in order to make the dataset own a balanced distribution in classes, this method randomly selects sample of minority class and copies the sample, until the number of samples in minority classes achieve the number of samples in majority class. Random sampling method replicates the known samples in order to make a balance dataset simply. However, it has not introduced additional information, resulting in a weak improvement in imbalanced classification.

SMOTE [6] method is a classic oversampling technique, this method selects a sample x from minority class dataset randomly, then randomly selects a sample y from k neighbors of the sample x, finally to generate a random number between 0 and 1, α, then artificially constructs a new sample for $g = x + \alpha \cdot (y - x)$, as shown in Fig. 3. Repeat this process until the dataset reaching balance. Although SMOTE method generates synthetic sample between the samples in minority class, and brings new information in dataset. It has improved the performance of imbalanced datasets classification, but the randomly sampling from the minority class means that all samples are equally important. Thus, SMOTE cannot reflects the difference of importance between samples.

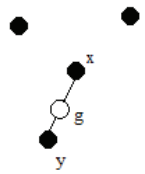

Fig. 3. Illustration of synthetic sample by SMOTE method

Borderline-SMOTE [7] is an improvement method of SMOTE. As the samples nearby classification boundary have higher information, the method considers to choose the samples close to classification boundary, and then generates synthetic samples between the sample and its neighbor by SMOTE method. Consequently, the synthetic sample has higher information as same, which are helpful to classifier. The key of Borderline-SMOTE is how to determine whether the sample of minority class is close to the classification boundary. Therefore, it establishes a k-NN graph on the dataset, and calculating the number of neighbors from majority class of each sample in minority class, δ, if $k/2<\delta<k$, the sample will be determined nearby the classification boundary. Once Borderline-SMOTE method determine the samples set close to classification boundaries, all the sample information in this set is the same.

ADASYN [8] method explicitly defines how much synthetic samples will be generate around the sample in minority class. The method use δ/k to calculate the proportion of neighbors from majority class of each samples in minority class. The larger of the proportion means the sample are closer to the classification boundary, and the sample has a higher information, it should be generate more synthetic samples around this sample.

Oversampling techniques like Borderline-SMOTE, ADASYN are usually applied to binary classification, because it is difficult to determine the classification boundary in multi-classes dataset. Therefore, it's difficult to migrate these methods in this situation. More imbalanced classification methods can be found in the literature [5].

Iterative Nearest Neighborhood Oversampling (INNO) [9] algorithm is a multi-classes imbalanced dataset processing method. This method is applied in the field of a semi-supervised learning, due to the quantity of labeled samples in semi-supervised learning areas is very few, but it contains a large amount of unlabeled samples. The method converts a certain number of unlabeled data to labeled data by the similarity between samples. However, the algorithm only use the distance as the similarity between samples, and has not considered the sample information.

Authorized licensed use limited to: UNIVERSITY TENAGA NASIONAL. Downloaded on May 14,2023 at 17:03:47 UTC from IEEE Xplore. Restrictions apply.

Based on problems in above methods, we propose a sample information measurement in multi-class imbalanced data classification, and apply it to INNO algorithm, to solve the drawback of losing sample information factors in the sampling process.

III. SAMPLE INFORMATION

A. The Application of Sample Information in Imbalanced Dataset Classification

For classification task, every sample in dataset has different contribution to classification result, it means some samples are more important and have more influence to classification result. To be simple, better representative samples could make better classification results. The importance of sample could be expressed by the sample information, in classification task, the sample information can be represented as the uncertainty of samples, higher uncertainty means harder to get the class of sample, also means it can give more contribution to classification result.

When use oversampling technique to deal with the problem of imbalanced dataset classification, the sample information decides what samples should be selected to generate synthetic sample around them. Higher sample information means its synthetic samples also have higher sample information, at the same time it can get better classification results when add these synthetic samples into the dataset. So to choose the sample which has higher sample information will have a significant influence on classification result.

B. Common Measurement of Sample Information

Several common measurements of imbalanced sample information are information entropy [10], margin sample [11], Best vs Second Best [12] etc.

In classification task, the sample information can be interpreted as uncertainty of classification for every sample, and expressed with formula $p(y_j|\ x_i)$. It represents that the probability of sample x_i belongs to class y_i. So the entropy of sample x_i can be expressed:

$$EP(x_i) = -\sum_{j=1}^{c} p(y_i \mid x_i) \log(\, p(y_i \mid x_i)) \qquad (1)$$

It can be seen that more entropy means higher uncertainty of classification for sample x_i.

In SVM classification method, the distance to classification boundary is considered as an effective approach to express the sample information. The closer to classification boundary means harder to identity the label of sample and the higher sample information it has. In Fig. 4, the distance to border of sample A, B, C, D is d_A、d_B、d_C、d_D, sample A and C are situated in hyper plane, so the distance to classification boundary is equal, $d_A = d_C$, but sample B and D are farer from the classification boundary. As result, it can be concluded that $d_B > d_D$, and $d_A > d_B$, so the sample information is $SI_A > SI_B > SI_C > SI_D$.

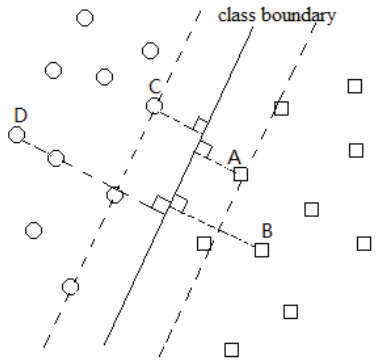

Fig. 4. the classification border of SVM

Acquirement of sample information above are depends on the classification result, or only suitable in binary-classification dataset. If the dataset is unbalance, the classification result becomes unreliable, or if the dataset has more than two classes, the measurements above cannot be facilitated.

Therefore, we consider getting sample information by the density of sample distribution without using classification method. The sample distribution in feature space is usually inhomogeneous. The samples in different region may have different sample information. Intuitively, for the sample lie in intensive region, its label has strong influence to their neighbors in classification result. The samples in sparse region have small influence in classification result to their neighbors in contrast. In Fig. 5, the region around sample A is more intensive than the region around sample B, so the label of sample A could influence more its neighbors. So we can use the density of region, which can also be described as local density, to describe the sample information.

Fig. 5. the schematic diagram of area density

Based on those considerations above, we proposed a sample information measurement based on local density. Local density also can be expressed by clustering methods [15], [16], but they need a high computational complexity.

C. Density-based Sample Information

Given a raw dataset $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$, $i \in [1, n]$, in which n represents the quantity of samples in dataset, and x_i is a vector of length m, which represents the *i-th* sample. Besides, m is the quantity of features for sample x_i, which can be expressed as $x_i = [x_{i1}, x_{i2}, \ldots, x_{im}] \in R^m$. In the dataset X, the labeled dataset is denoted as X_L, and the unlabeled dataset is denoted as X_U, and $X = \{X_L \cup X_U\}$, $|X_L| = l$, $|X_U| = u$, $l + u = n$. X_L can be represented as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_l, y_l)\}$, where $y_i \in \{1, 2, \ldots, c\}$ is the categories of the sample x_i and $i \in [1, l]$, in which c represents the quantity of categories.

Authorized licensed use limited to: UNIVERSITY TENAGA NASIONAL. Downloaded on May 14,2023 at 17:03:47 UTC from IEEE Xplore. Restrictions apply.

X_U can be represented as $\{x_{l+1}, x_{l+2}, \ldots, x_{l+u}\}$. The aim of classification is to obtain labels (categories or classes) of the samples in the unlabeled dataset.

In order to obtain the density of certain region, we establish the ε-NN graph on dataset at first. The method of establishing the ε-NN graph is to calculate the similarity of any two samples in the dataset X. If the similarity is bigger than the set value ε, an edge between x_i and x_j is established. The weight value of the edge represents the similarity between x_i and x_j which is represented as w_{ij}. Therefore, according to the similarity in the ε-NN graph, we can generate a similarity matrix $W = \{w_{ij}\}$, in which $w_{ij} = sim(x_i, x_j)$. For those values whose similarity is smaller than ε, we set them to 0. It can be represented as follows:

$$w_{ij} = \begin{cases} sim(x_i, x_j), & if\ sim(x_i, x_j) >= \varepsilon \\ 0, & else \end{cases} \quad (2)$$

As the samples in intensive region are more easily to establish edges with other samples, the sum of value in the corresponding matrix line is bigger. In order to obtain the information quantity of samples, we use the method of summing the similarity matrix lines. Thus, the information quantity SI_i of sample x_i can be represented as follows:

$$SI_i = \sum_j w_{ij} \quad (3)$$

After getting the sample information, we apply it to the sampling algorithm of the imbalanced datasets.

IV. SI-INNO

We utilize the INNO algorithm proposed in [9] to do the oversampling process on imbalanced datasets. At beginning, INNO converts some unlabeled data samples to labeled data samples only according to the similarities among them without their sample information. Such that, we propose an improved algorithm called SI-INNO with the sampling standard as equation 4.1 shows:

$$x_{maxk} = \arg\max_{x_k \subset X_U} SI_k \cdot sim(x_k, x_j) \quad (4)$$

We use the product of samples similarities and sample information to define the new "similarity" between samples. By this, the probability of the similar samples in the original similar matrix being chosen becomes different. Fig.6 depicts a dataset with four samples, and at beginning, the similarity between x_1 and x_2, as well as that between x_1 and x_3, are both 1. After the processing of new similarity standard, these two pair of similarities becomes 3 and 4 respectively. Since x_3 has more sample information, in the new similarity matrix, x_3 will get higher probability to be chosen.

$$W = \begin{matrix} x_1 & x_2 & x_3 & x_4 \end{matrix} \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{matrix} \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad SI = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 3 \end{bmatrix}, \quad SW = SI'.*W = \begin{bmatrix} 0 & 3 & 4 & 0 \\ 3 & 0 & 4 & 3 \\ 2 & 3 & 0 & 6 \\ 0 & 3 & 6 & 0 \end{bmatrix}$$

Fig. 6. New Similarity Matrix with SI-INNO

From the definition of sample information, we observe that the more sample information the data sample gets, the higher probability the sample to be chosen.

In INNO, the algorithm stops when the imbalanced degree of the dataset reaches the default level. However, since INNO is insensitive to the size of dataset, different datasets with different size may get the same imbalanced degree while they are different actually. In this paper, we propose an improved method called IR to describe the imbalanced degree according to standard deviation. We define the number of data samples of class j as r_j, and then we get a vector $r = \{r_1, r_2, \ldots, r_j, \ldots, r_c\}$, where $j \in [1, c]$. Such that, we use equation (4.2) to describe IR:

$$IR = \left(\frac{1}{c} \sum_{j=1}^{c} \left(\frac{r_j - \bar{r}}{l} \right)^2 \right), \quad \bar{r} = \frac{1}{c} \sum_{j=1}^{c} r_j \quad (5)$$

where l means the size of labeled dataset.

From Eq. 5, we can observe that the dataset is balanced when $IR = 0$, and it becomes imbalanced when $IR \neq 0$. By using the improved definition of IR, we can get a more rational description of imbalanced degrees of datasets.

SI-INNO algorithm can be depicted as Tab. 1.

TABLE I. PSEUDO-CODE OF SI-INNO

Algorithm 1 SI-INNO

Input: imbalanced labeled dataset X_L and unlabeled dataset X_U, neighbor number k, IR of dataset
Output: balanced or almost balanced dataset X_L', unlabeled dataset X_U'
Begin
1 Calculate similarity matrix W and sample information SI according to sample features
2 while $IR > 0$
3 find the minimum $r_j = \min\{r\}$ and its corresponding category c_j
4 initialize max = - ∞, maxk = 0
5 foreach labeled data x_j in c_j
6 foreach x_j neighbor x_k
7 if x_k in L or x_k is a neighbor of other labeled data
8 continue
9 end if
10 if SI_k * w_{jk} > max
11 update max，maxk
12 end if
13 end foreach
14 end foreach
15 if maxk = 0 // all neighbors of labeled data of category c_j has neighbor relationship with labeled data in other categories
16 $r_j = \max\{r\}$, continue
17 end if
18 mark x_{maxk} the category j, remove it from X_U, and add to X_L
19 $r_j = r_j + 1$, update IR
20 end while

V. EXPERIMENT

A. Dataset description and experiment settings

In order to reflect the accuracy of the classifier after the SI-INNO algorithm deals with the imbalanced dataset accurately, we will set up experiments that are under sets of different imbalances on each dataset, and different imbalance will be produced by collecting a different number of labeled samples randomly from datasets, and the experiment of the same quantity of labeled dataset will be carried out for several

Authorized licensed use limited to: UNIVERSITY TENAGA NASIONAL. Downloaded on May 14,2023 at 17:03:47 UTC from IEEE Xplore. Restrictions apply.

times to avoid the impact caused by a particular experiment. In practice, the experiment of the same imbalance will take the average of 50 cycles.

All datasets use RBF kernel function to measure the similarity between samples, and SI-INNO parameters' similarity parameter ε, number of neighbor k, stop condition s and RBF parameter σ will be selected an appropriate value based on the actual dataset. In the experiment, we will use three different datasets, the description of datasets is showed in Tab. II.

TABLE II. DATASET DESCRIPTION

dataset	number	class	feature	proportion
IRIS	150	3	5	50:50:50*
WINE	178	3	13	59:70:47
GLASS	214	6	10	70:76:17:13:9:29

* construct imbalance artificially

On each dataset, we will artificially construct the imbalance of datasets. On IRIS dataset, we set the number of labeled samples in different class to a different quantity. The quantity of samples in "setosa" category varies from [1,10], "virginica" category varies from [10,19], and "versicolor" category is set to be 10. On WINE dataset, we set the number of labeled samples in category "1" varied from [10,1], category "2" varied from [10,19], and category "3" is set to be 10. On GLASS dataset, we set the number of labeled samples in category "1", "5", "6" varied from [2,10], and the other categories are set to be 2.

B. The analysis of results

In order to better explain that SI-INNO algorithm outperform INNO algorithm on imbalanced datasets classification. Similar to INNO algorithm, we combine SI-INNO-sampling method with two classical semi-supervised learning GRF [13], LGC [14] method to classify the datasets. In comparison, we use INNO + GRF and INNO + LGC from INNO's paper, and classification results on IRIS, WINE, GLASS dataset are respectively showed in Fig. 7, Fig. 8 and Fig. 9.

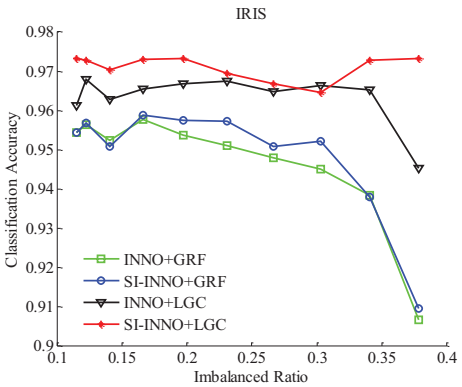

Fig. 7. Classification accuracy on GLASS dataset

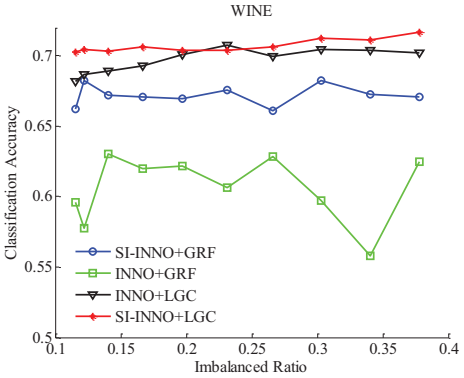

Fig. 8. Classification accuracy on WINE dataset

We can see it from Fig.7, when imbalance of IRIS dataset increases, SI-INNO + LGC, SI-INNO + GRF algorithm classification accuracy is better than INNO + LGC, INNO + GRF algorithm, and the classification accuracy of SI-INNO + LGC algorithm is more stable, which is 97%-98%, better than the classification accuracy of the INNO + LGC algorithm, which is 96% -97%. While the imbalance of dataset is about 0.4, except SI-INNO + LGC algorithm, classification accuracy of other algorithms declines more obviously. Therefore, the combination of the sample information of SI-INNO data resampling methods for classification of datasets improves the imbalanced classification more obviously.

We can also see it from Fig.8 and Fig. 9 that the algorithm SI-INNO + LGC and SI-INNO + GRF combining with sample information perform a better robustness on each dataset, and classification accuracy is higher, but the classification accuracy of the contrast algorithm is poor. Therefore, we can conclude SI-INNO algorithm combining with sample information shows a better sampling result for multi-classes imbalanced datasets.

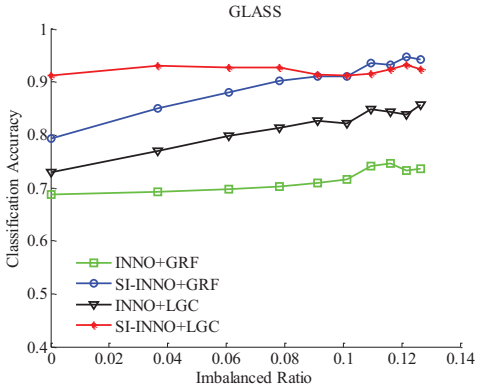

Fig. 9. Classification accuracy on GLASS dataset

CONCLUSION

Sample information plays an important role in the process of the sampling on imbalance datasets. It determines what samples should be sampled to make advantage for imbalanced datasets classification. We proposed a sample information measurement based on local density to determine what

Authorized licensed use limited to: UNIVERSITY TENAGA NASIONAL. Downloaded on May 14,2023 at 17:03:47 UTC from IEEE Xplore. Restrictions apply.

samples have more contribution to classification. Therefore, we apply the proposed sample information method with a multi-classes imbalanced classification algorithm INNO to make its sampling criterion become more reasonable. Experiments on several multi-class imbalanced datasets show that the algorithms combining with the sample information perform better classification accuracy. Thus, the proposed method indeed improves the sampling result effectively on imbalanced datasets.

REFERENCES

[1] Wang S. "A comprehensive survey of data mining-based accounting-fraud detection research," Proceedings of International Conference on Intelligent Computation Technology and Automation (ICICTA), Changsha, China, 2010, pp. 50-53.

[2] Xu G, Niu Z, Gao X, et al. "Imbalanced text classification on host pathogen protein-protein interaction documents," Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE), Singapore, Singapore, ,2010, pp. 418-422.

[3] Li F, Li G, Yang N, et al. Label matrix normalization for semisupervised learning from imbalanced Data. New Review of Hypermedia and Multimedia, 2014, vol. 20, pp. 5-23.

[4] Estabrooks A, Jo T, Japkowicz N. "A multiple resampling method for learning from imbalanced data sets," Computational Intelligence, 2004, vol. 20, pp. 18-36.

[5] He H, Garcia E A. "Learning from Imbalanced Data," IEEE Transactions on Knowledge And Data Engineering, 2009, vol. 21,pp. 1263-1284.

[6] Chawla N V, Bowyer K W, Hall L O, et al. "SMOTE: Synthetic minority over-sampling technique,", Journal Of Artificial Intelligence Research, 2002, vol. 16, pp. 321-357.

[7] Han H, Wang W Y, Mao B H. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," Advances in intelligent computing. Springer Berlin Heidelberg, 2005, pp. 878-887.

[8] [He H, Bai Y, Garcia E A, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," Proceedings of IEEE International Joint Conference on Neural Networks, Hong Kong, China, 2008, pp. 1322-1328.

[9] Li F, Yu C, Yang N, et al. "Iterative nearest neighborhood oversampling in semisupervised learning from imbalanced data." The Scientific World Journal, 2013.

[10] Coifman R, Wickerhauser M. "Entropy-based algorithms for best basis selection," IEEE Transactions on Information Theory, 2002, vol. 38, pp. 713-718.

[11] Ertekin S, Huang J, Giles C L. "Active learning for class imbalance problem," Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007, pp. 823-824.

[12] Joshi A J, Porikli F, Papanikolopoulos N. "Multi-class active learning for image classification," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, US, 2009, pp. 2372-2379.

[13] Zhu X, Ghahramani Z, Lafferty J. "Semi-supervised learning using Gaussian fields and harmonic functions," Proceedings of the Twentieth International Conference on Machine Learning, Washington DC, US, 2003, pp. 912-919.

[14] Zhou D, Bousquet O, Lal T N, et al. "Learning with local and global consistency," Advances in neural information processing systems, 2004, vol.16, pp. 321-328.

[15] Q. Zhang, and Z. Chen, "A Weighted Kernel Possibilistic C-means Algorithm Based on Cloud Computing for Clustering Big Data," International Journal of Communication Systems, vol.27, no.9, pp.1378-1391, 2014.

[16] Q. Zhang and Z. Chen, "A High-order Possibilistic C-means Algorithm for Clustering Incomplete Multimedia Data", IEEE Sy stems Journal, 2015. DOI: 10.1109/JSYST.2015.2423499.

Authorized licensed use limited to: UNIVERSITY TENAGA NASIONAL. Downloaded on May 14,2023 at 17:03:47 UTC from IEEE Xplore. Restrictions apply.