Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print), ISSN:2521-3520 (online)   Volume 31, Issue 2, 2020   DOI: http://doi.org/10.23851/mjs.v31i2.740

Research Article   Open Access

Adaptation Proposed Methods for Handling Imbalanced Datasets based on Over-Sampling Technique

Liqaa M. Shoohi, Jamila H. Saud*

Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, IRAQ.

*Correspondent author email: dr.jameelahharbi@gmail.com, liqaa.m.shoohi@gmail.com

ArticleInfo

Received
06/09/2019

Accepted
29/11/2019

Published
15/04/2020

ABSTRACT

Classification of imbalanced data is an important issue. Many algorithms have been developed for classification, such as Back Propagation (BP) neural networks, decision tree, Bayesian networks etc., and have been used repeatedly in many fields. These algorithms speak of the problem of imbalanced data, where there are situations that belong to more classes than others. Imbalanced data result in poor performance and bias to a class without other classes. In this paper, we proposed three techniques based on the Over-Sampling (O.S.) technique for processing imbalanced dataset and redistributing it and converting it into balanced dataset. These techniques are (Improved Synthetic Minority Over-Sampling Technique (Improved SMOTE), Borderline-SMOTE + Imbalanced Ratio(IR), Adaptive Synthetic Sampling (ADASYN) +IR) Algorithm, where the work these techniques are generate the synthetic samples for the minority class to achieve balance between minority and majority classes and then calculate the IR between classes of minority and majority. Experimental results show ImprovedSMOTE algorithm outperform the Borderline-SMOTE + IR and ADASYN + IR algorithms because it achieves a high balance between minority and majority classes.

KEYWORDS: Imbalanced Datasets; Over-Sampling; SMOTE; Borderline-SMOTE; ADASYN.

الخلاصة

تم التعرف على مشكلة عدم التوازن الطبقي كمشكلة بحثية مهمة في التصنيف في السنوات الأخيرة ، حيث أدخلت عددا من الأساليب لتحسين دقة التصنيف باستخدام عدد من الطرق لإعادة توازن توزيعات الفصل مثل : (أخذ عينات أقل في التعلم أو أخذ العينات الزائدة لتعلم مجموعات البيانات) والتي تعطي أداءً جيدًا.تم تطوير العديد من الخوارزميات للتصنيف ، مثل الشبكات العصبية ذات الانتشار الخلفي (BP)، وشجرة القرارات ، وشبكات بايزي ، وما إلى ذلك ، وقد تم استخدامها مرارًا وتكرارًا في العديد من المجالات. تتحدث هذه الخوارزميات عن مشكلة البيانات غير المتوازنة ، حيث توجد عينات تنتمي إلى فئات أكثر من غيرها. تؤدي البيانات غير المتوازنة إلى ضعف الأداء والانحياز لفئة دون فئات أخرى.في هذه الورقة اقترحنا ثلاثة تقنيات لمعالجة مجموعة البيانات غير التوازنة واعادة توزيعها وتحويلها الى بيانات متوازنة وهذه التقنيات هي: (Improved SMOTE, Borderline-SMOTE+IR, ADASYN +IR) Algorithm حيث ان عمل هذه التقنيات هو توليد عينات اصطناعية لفئة الأقلية لتحقيق التوازن بين فئات الاقلية والاغلبية وبعدها يتم حساب النسبة غير المتوازنة (IR) بين فئات الأقلية والأغلبية. تظهر النتائج التجريبية أن خوارزمية ImprovedSMOTE تتفوق على خوارزميات Borderline-SMOTE + IR و ADASYN + IR لأنها تحقق توازنًا كبيرًا بين فئات الأقلية والأغلبية.

INTRODUCTION

The data is imbalanced when the data distribution is not systematic across different classes [1]. In many applications learning occurs with the distribution of class imbalances regularly, and this situation occurs in the data when the number of examples in the minority class is much lower than number of examples in class of the majority .This means that the number of examples into the classes exceeds the number of examples in other classes [2], the class which contains a large number is called negative or majority class, either for the class which contains a few numbers of examples called positive or minority class. Minority class is a great impact when misclassified that will be considered as interest class [3]. it is likely that the examples of minority class may be ignored or external values led to poor classification compared to the majority, the dominant class [4].

Many problems of real-world contains several concepts with very few examples in a large group and are also described the cost of obtaining them or through their scarcity. Categories are either two classes or multi-class. Required analyze methods to address the problem of multi-class data imbalance which focuses on many real-world problems such as medical diagnosis, fraud detection, defect detection software, network intrusion [5][6]. There are simple ways to overcome the problem of class imbalance in typical learning. It is usually to test the training data by taking examples of the minority too much. Another approach is cost-sensitive learning which in turn rewrites the algorithms by weighing the examples of the minority class [7].

Techniques are used to process the imbalanced data classification such as [8]:

- ***Approach of Data Level***

Use imbalanced native dataset to obtain balanced dataset by using algorithms of machine learning to obtain desired results.

- ***Approach of Algorithm Level***

Use algorithms that can handle unbalanced of data.

- ***Approach of Cost-Sensitive***

Hybrid technology of the first and the second techniques are combined to achieve a decrease in the costs of poor classification and accuracy.

Data Level Approach is classified into different sets

1. Over-Sampling (O.S.) Technique

O.S. technique, the data are balanced by adding examples of the minority class to the original data. Figure1 shows how to generate synthetic samples and add them to the original data by using Synthetic Minority Over-Sampling Technique (SMOTE) algorithm [9][10].

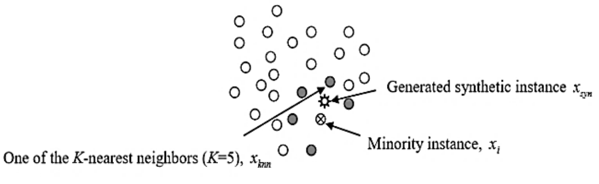

Generated synthetic instance x_{syn}
Minority instance, x_i
One of the K-nearest neighbors (K=5), x_{knn}

Figure 1. Synthetic Oversampling example by SMOTE algorithm.

2. Undersampling Technique

Undersampling technique, the data is balanced by removing examples from the majority class of the original data.

3. Hybrid Technique

Hybrid technique, the data is balanced by combining previous techniques, first O.S. technique is used and then undersampling technique is used [3].

Major brain tumors occur in around 250,000 people annually at the international level, accounting for less than 2% of diseases. In more than 15 young people, brain tumors are second only to intensive lymphocytic leukemia as a cause of malignancy [4].

AIM OF RESEARCH

In order to predict of all classes the majority and the minority, in other words, the minority class is also classified. The imbalanced dataset must be converted into a balanced dataset. This is done by applying the approach proposed in our thesis that will improve the performance of the classifier with accurate values.

LITERATURE REVIEW

Many research has been completed to solution the imbalance in O.S. datasets. Tanghui et al [9] in their research discusses the predicted user's quality of experience (QoE) assessment of IPTV which is the prediction of the user's complaint call. The status data was collected from set-top box with data for the user's complainant, and the suitable QoE user model prediction was selected. First, based on obtained imbalanced dataset from Jiangsu Telecom, where on majority class has be applying randomization under-sampling after data cleaned and a feature selection for data. K-means was then proposed with the SMOTE algorithm to samples generate for the minority class. Also was built Naïve Bayes model. In their works the results experimental displayed which improved schema could improve predictive accuracy and as well show that the SMOTE integrated algorithm with K-means implements the user’s complaints in predicting best than the Borderline-SMOTE algorithm. Haibo He et al., [11] are proposed the ADASYN method to facilitate learning for imbalanced dataset. ADASYN method is used to achieve two objectives: i. Adaptive learning ii. Minimize bias, and this algorithm addresses the problem of classification with two class. Simulation results the effectiveness of this method is shown on five the datasets based on different the evaluation measures. Search for Path Planning” In this paper, both the cuckoo search and the bat algorithm are linked to the proposed problem and seen in simulation results. The systems are associated with a number of populations, and the bat algorithm gives better results when compared to the cuckoo search.

PROPOSED METHODS

The proposed adaptation methods in this paper consists of the following:

A. ImprovedSMOTE Algorithm

After reading dataset and determining the majority and minority classes, we will have suggested many steps to Improved SMOTE algorithm such as:

Step1:Removes the noise from the minority class.

Step2:Apply clustering on minority class, divide the minority class into four clusters and then select the appropriate cluster to generate the synthetic samples, i.e, the appropriate cluster, is the cluster that achieves a higher imbalanced ratio between the majority and minority classes.

Step3:SMOTE algorithm is used to generate synthetic samples [9].

Step4:Using imbalanced ratio (IR) function:

$$IR = m_s / m_j \quad \ldots\ldots\ldots\ldots (1)$$

Where IR is imbalanced ratio, $\boldsymbol{m_s}$ is number of the minority class examples, $\boldsymbol{m_j}$ is number of the majority class examples.

Figure 2 shows the structure of our adaptation of SMOTE algorithm to Improved.

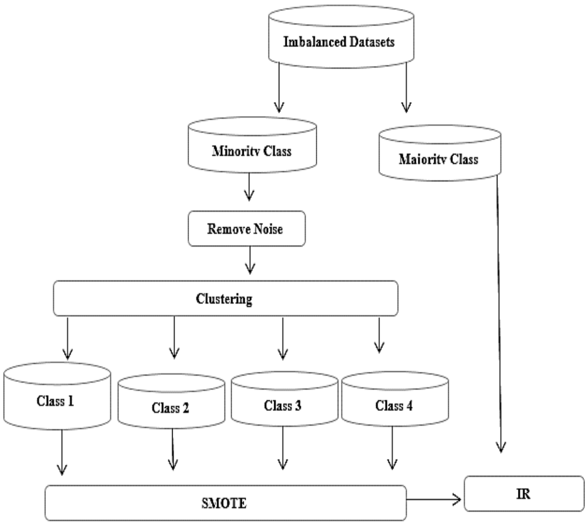


Figure 2. Structure of ImprovedSMOTE algorithm.

B. Adaptive Borderline-SMOTE+IR Algorithm

After reading dataset, the adaptive Borderline-SMOTE+IR algorithm consists of three steps:

Step1:The Borderline-SMOTE algorithm is used to generate synthetic samples [9].

Step2:Determine the output dataset of Borderline-SMOTE algorithm into majority class and minority class.

Step3: Use equation (1).

Figure 3 shows the structure of Borderline-SMOTE+IR algorithm.

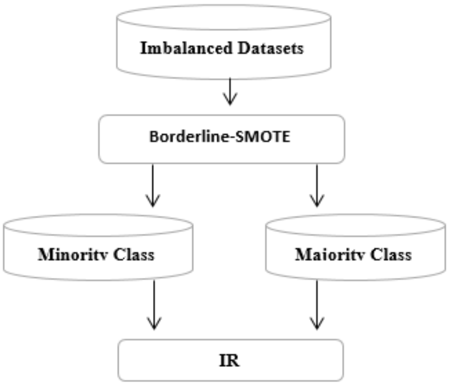


Figure 3. Structure of Borderline-SMOTE+IR Algorithm.

C. Adaptive ADASYN +IR Algorithm

After reading dataset, the adaptive ADASYN +IR algorithm consists of three steps:

Step1: Divide the dataset into majority class and minority class.

Step2: Apply The ADASYN algorithm to generate synthetic samples [11].

Step3: Use equation (1).

Figure 4 shows the structure of ADASYN +IR algorithm.

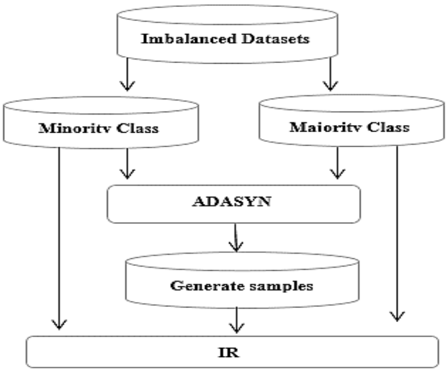


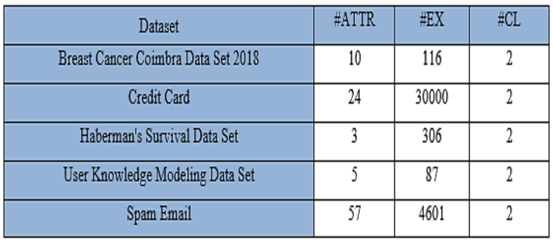
Figure 4. Structure of ADASYN +IR Algorithm.

EXPERIMENTAL RESULTS OF ADAPTATION IMBALANCED DATASET

A. Datasets

We have applied O.S. technique to datasets (binary class) obtained from the University of California Irvine (UCI) machine learning repository and Knowledge Extraction based on Evolutionary Learning (KEEL) database repository, which differ in the class of imbalances and also vary in size, features and numbers. Table 1 shows the properties of the required datasets, including the name of the dataset, the number of instance (#EX), the number of features (#ATTR) and the number of classes (#CL) [3].

Table 1. Characteristics of Dataset



Dataset	#ATTR	#EX	#CL
Breast Cancer Coimbra Data Set 2018	10	116	2
Credit Card	24	30000	2
Haberman's Survival Data Set	3	306	2
User Knowledge Modeling Data Set	5	87	2
Spam Email	57	4601	2

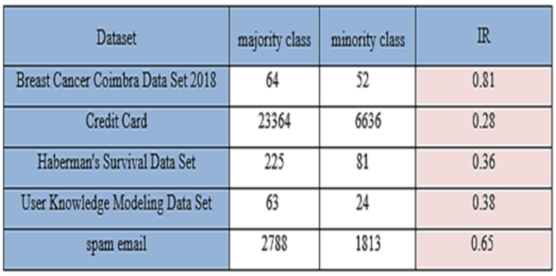
B. Experimental Setup

In all of our experiments IR is used as the basic criterion for evaluating the algorithm. When the IR is close to number 1, this means that the number of instance in the majority class is close to the number of instance in the minority class. Thus the balance between classes is achieved.

C. The Results of Adaptation Methods

Table 2 shows the division of the dataset into a majority and a minority classes, and then calculates the IR between them.

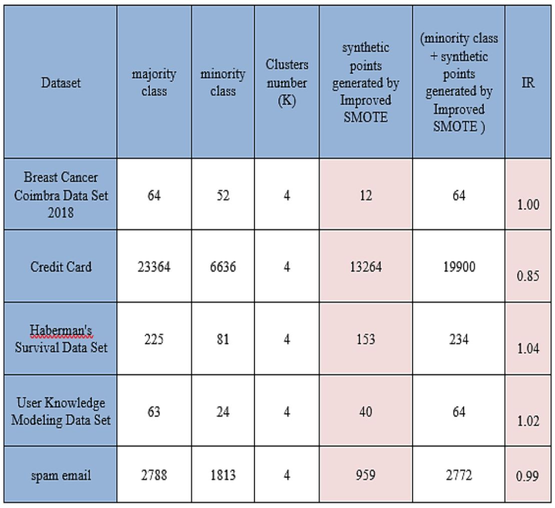
Table 2. Result of the identification of the majority and minority classes and calculate the IR of original datasets.



Dataset	majority class	minority class	IR
Breast Cancer Coimbra Data Set 2018	64	52	0.81
Credit Card	23364	6636	0.28
Haberman's Survival Data Set	225	81	0.36
User Knowledge Modeling Data Set	63	24	0.38
spam email	2788	1813	0.65

Table 3 shows the synthetic samples generated by SMOTE and the IR account between (minority class + synthetic points generated by Improved SMOTE) and majority class.

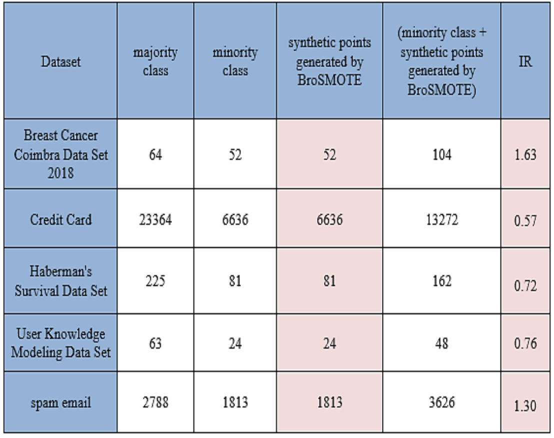
Table 3. Result of the Improved SMOTE Algorithm.



Dataset	majority class	minority class	Clusters number (K)	synthetic points generated by Improved SMOTE	(minority class + synthetic points generated by Improved SMOTE)	IR
Breast Cancer Coimbra Data Set 2018	64	52	4	12	64	1.00
Credit Card	23364	6636	4	13264	19900	0.85
Haberman's Survival Data Set	225	81	4	153	234	1.04
User Knowledge Modeling Data Set	63	24	4	40	64	1.02
spam email	2788	1813	4	959	2772	0.99

Table 4 shows the synthetic samples generated by Borderline-SMOTE and the IR account between (minority class + synthetic points generated by Borderline-SMOTE) and majority class.

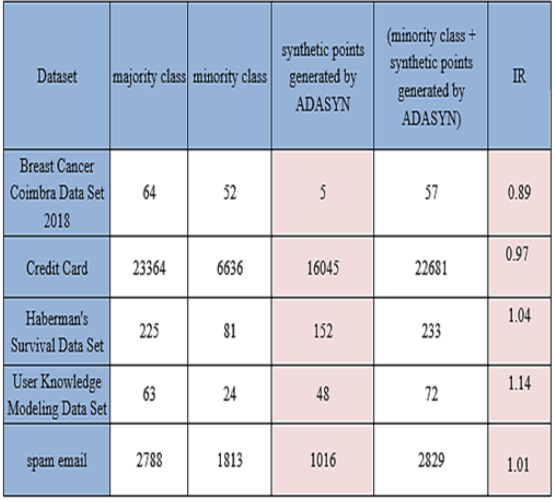
Table 4. Result of the Borderline-SMOTE+IR Algorithm.



Dataset	majority class	minority class	synthetic points generated by BroSMOTE	(minority class + synthetic points generated by BroSMOTE)	IR
Breast Cancer Coimbra Data Set 2018	64	52	52	104	1.63
Credit Card	23364	6636	6636	13272	0.57
Haberman's Survival Data Set	225	81	81	162	0.72
User Knowledge Modeling Data Set	63	24	24	48	0.76
spam email	2788	1813	1813	3626	1.30

Table 5 shows the synthetic samples generated by ADASYN and the IR account between (minority class + synthetic points generated by ADASYN) and majority class.

Table 5. Result of the ADASYN +IR Algorithm.



Dataset	majority class	minority class	synthetic points generated by ADASYN	(minority class + synthetic points generated by ADASYN)	IR
Breast Cancer Coimbra Data Set 2018	64	52	5	57	0.89
Credit Card	23364	6636	16045	22681	0.97
Haberman's Survival Data Set	225	81	152	233	1.04
User Knowledge Modeling Data Set	63	24	48	72	1.14
spam email	2788	1813	1016	2829	1.01

CONCLUSIONS

In this paper, The proposed adaptation methods based on O.S. technique Which will convert the imbalanced dataset into balanced dataset because it gives the number of generated images that achieve balance between classes of majority and minority. When implemented are: ImprovedSMOTE, Borderline-SMOTE+IR, and ADASYN+IR algorithms shows that the results of the experiment of ImprovedSMOTE algorithm outperform the Borderline-SMOTE+IR and ADASYN+IR algorithms by converting the imbalanced dataset into a balanced dataset ,because the ImprovedSMOTE algorithm achieves a high IR between classes minority and majority from Borderline-SMOTE+IR and ADASYN+IR algorithms when applied to datasets with different IR.

FUTURE WORKS

The implementation of these techniques in a system and adjusted according to the need in the application to diagnose diseases.

REFERENCES

[1] Y. Yan, “Deep Learning Based Imbalanced Data Classification and Information Retrieval for Multimedia Big Data,” ProQuest Diss. Thesis, p. 172, 2018.

[2] N. V Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, “SMOTEBoost : Improving Prediction,” Lavrač N., Gamberger D., Todorovski L., Blockeel H. Knowl. Discov. Databases PKDD 2003. LNCS, vol. 2838, pp. 107–119, 2003.

[3] W. Pedrycz and S. Chen, “Data Science and Big Data: An Environment of Computational Intelligence,” Springer International Publishing AG 2017, vol. 24. 2017.

[4] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: A review,” Int. J. Adv. Soft Comput. its Appl., vol. 7, no. 3, pp. 176–204, 2015.

[5] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, “Safe-Level-SMOTE : Safe-Level-Synthetic Minority Over-Sampling Technique,” Springer-Verlag Berlin Heidelberg 2009, pp. 475–476, 2009.

[6] B. J. Park, S. K. Oh, and W. Pedrycz, “The Design of Polynomial Function-based Neural Network Predictors for Detection of Software Defects,” Inf. Sci. (Ny)., vol. 229, pp. 40–57, 2013.

[7] Qi Dong, Shaogang Gong, and Xiatian Zhu, “Imbalanced Deep Learning by Minority Class Incremental Rectification,” arXiv:1804.10851v1, 28 Apr 2018

[8] S. Del Río, V. López, J. M. Benítez, and F. Herrera, “On the Use of MapReduce for Imbalanced Big Data using Random Forest,” Inf. Sci. (Ny)., vol. 285, no. 1, pp. 112–137, 2014.

[9] T. Wang, R. Huang, X. Wei, and F. Zhou, “Improving User’s Quality of Experience in Imbalanced Dataset,” Proc. - 2016 Int. Comput. Symp. ICS 2016, pp. 690–695, 2017.

[10] J. M. Choi, “A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines,” AAAI’2000 Work. imbalanced datasets, p. 107, 2010.

[11] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” 978-1-4244-1821-3/08 IEEE, 2008.