Data Science in Finance and Economics

http://www.aimspress.com/journal/dsfe

DSFE, 1(1): 21–36.
DOI: 10.3934/DSFE.2021002
Received: 01 May 2021
Accepted: 25 May 2021
Published: 01 June 2021

Research article

Learning from class-imbalanced data: review of data driven methods and algorithm driven methods

Cui Yin Huang and Hong Liang Dai*

School of Economics and Statistics, Guangzhou University, Guangzhou 510006, China

* **Correspondence:** Email: hldai618@gzhu.edu.cn.

Abstract: As an important part of machine learning, classification learning has been applied in many practical fields. It is valuable that to discuss class imbalance learning in several fields. In this research, we provide a review of class imbalanced learning methods from the data driven methods and algorithm driven methods based on numerous published papers which studied class imbalance learning. The preliminary analysis shows that class imbalanced learning methods mainly are applied both management and engineering fields. Firstly, we analyze and then summarize resampling methods that are used in different stages. Secondly, we provide a detailed instruction on different algorithms, and then we compare the results of decision tree classifiers based on resampling and empirical cost sensitivity. Finally, some suggestions from the reviewed papers are incorporated with our experiences and judgments to offer further research directions for the class imbalanced learning fields.

Keywords: class imbalanced learning; classification; machine learning; data mining; deep learning

JEL Codes: TP39

1. Introduction

The class imbalance problem refers to the hot potato that the quantity of one class presents abnormal characteristic, which is much larger or less than the other classes of samples and the cost of misclassification between this classes of samples is different, leading to failure for standard classifiers. Thus, characteristics of class imbalanced datasets are shown as follows: the quantity imbalanced of different classes of samples and the cost imbalanced of miscalculation (Li et al., 2019). Usually, class imbalanced learning methods are considered as the technologies that can solve the above problem, which are widely used in several files such as bioinformatics (Blagus and Lusa, 2013), software defect monitoring (Lin and Lu, 2021), text classification (Ogura et al., 2011), and computer vision (Pouyanfar and Chen, 2015) etc. Therefore, these broad applications reveal tremendous value to research class imbalanced learning methods.

Standard classifiers such as logistic regression (LR), Support Vector Machine (SVM) and decision tree (DT) are suitable for balanced training sets. When facing imbalanced scenarios, these models often provide suboptimal classification results (Ye et al., 2019). For example, when facing imbalanced datasets, it is possible that unsatisfactory classification result was produced by Bayesian classifier, and the unsatisfactory classification result was influenced by the overlapping range of different class in the sample space (Domingos and Pazzani, 1997). Similarly, when the SVM classifier is employed to handle class imbalanced datasets, the optimal hyperplane will move to the core range of the majority class. Particularly, when data sets present the characteristic of highly imbalanced (Jiang et al., 2019) or interclass aggregation (Zhai et al., 2010), we obtained outcome that all sub-cluster samples of the minority class will be misclassified.

Therefore, the class imbalanced influenced the result of standard classifiers (Yu et al., 2019). Generally speaking, the class imbalance ratio (IR) is defined as the ratio of majority class size to minority class size, which can measure the degree of class imbalance in data sets. According to literature analysis, the result of standard classifiers influenced by class imbalanced was generally positive proportion that the greater IR was, the greater the impact has (Cmv and Jie, 2018). However， we should realize that class imbalance does not always lead poor results to the classifier. In addition, the following are also some factors that affect the results of standard classifiers:

- The scale of the overlapping space, which refers to the feature that different classes of samples have no clear boundary in the sample space.
- The number of noise samples, which refers to a few examples of one class far away from the core area of the class (López et al., 2015).
- The number of training samples, which refers to the model training samples (Yu et al., 2016).
- The degree of interclass aggregation, which refers to this feature that one class samples present two or more clusters in the sample space， and these clusters can distinguish major and minor (Japkowicz et al., 2002).
- The dimension of dataset, which refers to the number of features.

The above factors lead a suboptimal result. It is worth noting that when the above factors appear in the imbalanced datasets, worse results will emerge than of in the balanced scenario. Here, we generated a series of data sets to verify the influence of these factors on standard classifiers. Detailed results are shown in the appendix.

In this research, we aim to provide an overview of class imbalanced learning methods. The rest of this research is organized as follows. Section 2 introduces approaches to addressing class imbalanced dataset both data driven and algorithm driven. Section 3 provides a review of measurement of classifier performance to class imbalanced classifiers. In Section 4, we discuss our opinions for the challenges and directions of future from to analysis of relevant literature. Finally, Section 5 presents the conclusions of this study.

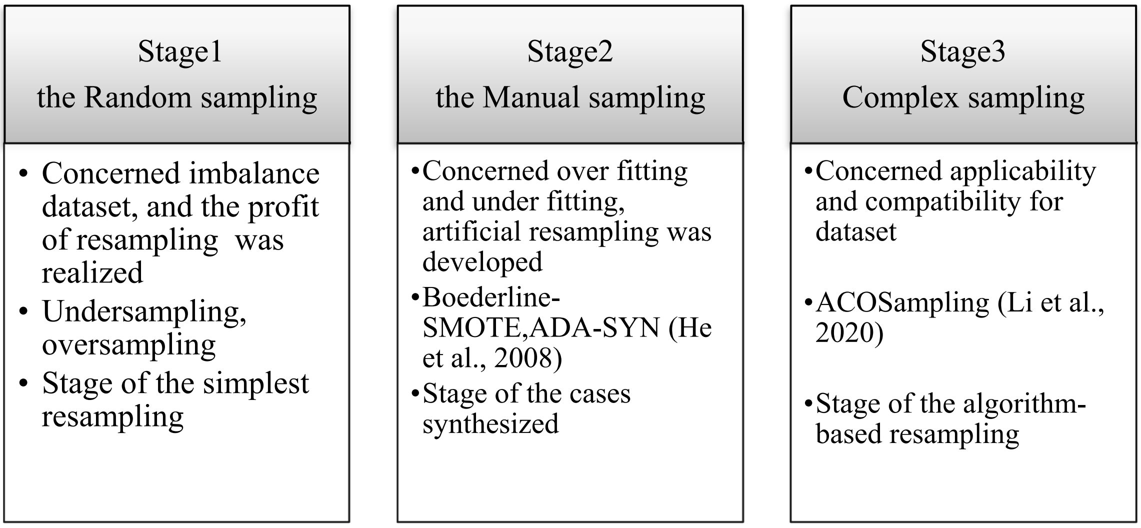
2. Data driven methods and algorithm driven methods

The research of class imbalance learning originated in the late 1990s. Since then, numerous methods have been developed. Thus, this study discusses the key methods to handle class imbalanced problems from data driven (Liu et al., 2019) and algorithm driven (Wu et al., 2019).

2.1. Data driven methods

Methods from the data drive, also known as data-level methods or resampling methods. These methods reverse the property of the imbalance characteristics of classes' quantity by randomly generating cases of the minority class (ROS) or removing cases of the majority class (RUS). It can be regarded as one of data preprocessing processes, therefore, resampling and the classifier training processes are independent on each other, and it was compatible with standard classifiers (Maurya and Toshniwal, 2018; Wang and Minku, 2015).

About methods of data driven can be described as follows.



Stage1
the Random sampling

- Concerned imbalance dataset, and the profit of resampling was realized
- Undersampling, oversampling
- Stage of the simplest resampling

Stage2
the Manual sampling

- Concerned over fitting and under fitting, artificial resampling was developed
- Boederline-SMOTE,ADA-SYN (He et al., 2008)
- Stage of the cases synthesized

Stage3
Complex sampling

- Concerned applicability and compatibility for dataset
- ACOSampling (Li et al., 2020)
- Stage of the algorithm-based resampling

Figure 1. The three stages of methods from data driven.

Table 1. The summary of some methods from data driven.

	Methods and illustrations
Over-sampling	ROS: generated the cases of the minority class randomly
	SMOTE: generated the cases of the minority class with KNN randomly
	Borderline-SMOTE: generated the cases of the minority with SMOTE in overlapping range
	EOS: generated the cases of the minority class randomly with "entropy" information
	RBO: generated the cases of the minority class randomly with "radial" information
Under-sampling	RUS: removed the cases of the majority class randomly
	SMOTE + ENN (Tao et al., 2019): removed the cases of the majority class with KNN randomly
	SMOTE + Tomek (Wang et al., 2019): removed the cases of the majority with deleting Tomek cases
	OSS (Rodriguez et al., 2013): removed the cases of the majority with just deleting the case of the majority in Tomek cases
	SBC (Xiao and Gao, 2019): removed the cases of the majority class with the clustering theory randomly
	EUS: removed the cases of the majority class randomly with "entropy" information
Hybrid sampling	EHS: hybrid resampling that entropy based
	CCR: hybrid resampling that synthesizing and cleaning based

Firstly, researchers pointed out that the random resampling can be used to deal with class imbalanced datasets, which was the simplest data-driven method to improve the classification accuracy of the minority class. But the uncomplicated data driven methods present some shortcomings, for instance, longer learning time, more running memory and poor generalization ability were presented by Oversampling due to the repeatability of samples. In addition, Undersampling will reduce the performance of classification owing to the lack of information resulted from the elimination of samples. Secondly, as the disadvantages of simple random sampling technology were exposed, some better methods were developed such as the synthetic minority oversampling techniques (SMOTE) (Chawla et al., 2011) and Borderline-SMOTE (Hui et al., 2005). The former method was proposed by Chawla et al. (2002). It was an oversampling algorithm that based on k-nearest neighbor (KNN) to synthesize a new virtual sample of the minority classes randomly among the minority class. Compared with ROS, SMOTE had a stronger ability of generalization and overcame overfitting in a certain extent. Borderline-SMOTE was an oversampling strategy based on SMOTE. Borderline-SMOTE synthesizes mainly the minority class samples at the class boundary, therefore, the method's classification result was better than SMOTE when one dataset with a few noise samples. In recent years, with the continuous progress of computer technology, some more superior methods have been proposed such as the cleaning resampling method (Koziarski et al., 2020), and based-radial undersampling method (Krawczyk et al., 2020), etc. Analyzing data driven methods, Yu deemed that the data driven methods underwent three stages random sampling technology stage, manual sampling technology stage and complex algorithm stage (Yu, 2016), shown in Figure 1.

Table 1 provides a summary of data driven method and its analysis. We can draw that the overlapping was important factor affecting the impact of standard classifiers, and that different cases were have different impacts on classification in the sample space, and that some concepts were defined by researchers such as “energy (Li L et al., 2020)” to provide some information for resampling.

2.2. Algorithm driven methods

Data driven methods are regarded as independent of classifiers methods. Yet algorithm driven methods are regarded as dependent classifiers methods. These methods are improving standard classifiers that include cost-sensitive learning based and threshold moved based mainly. For methods of algorithm driven, the main core algorithm was cost sensitive learning, and the supported learning algorithms include four learning technologies: active learning, decision compensation learning, feature extraction learning and emblems learning.

2.2.1. Cost-sensitive learning

Cost sensitive learning was one of the frequently used technologies to solve the problem of class imbalanced (Wan and Yang, 2020), and its goal is to minimize the cost of overall misclassification. In the process of model learning, according to the practical problems, different factors of penalty cost were given to different classes. Cost sensitive learning’s core is the design of cost matrix which could combine with the standard classifiers model to improve classification result (Zhou and Liu, 2010). For instance, we could obtain a posteriori probability which was more suitable for dealing with class imbalance problems, though the original Bayesian classifier posterior probability was fused with cost matrix (Kuang et al., 2019). And DT classifier integrated cost matrix into the process of attribute selection and pruning for the purpose of optimizing the classification result (Ping et al., 2020).

What the above analysis shows is that the technology was strongly dependent on cost matrix. Main design methods are as follows:

- Empirical weighted design, which shows that the cost coefficients of the samples of the same class are the same (Zong et al., 2013).
- Fuzzy weighted design, which shows that the cost coefficients of the same class are different in different position of sample spaces (Dai, 2015).
- Adaptive weighted design, which is iterative and dynamic, converging to the global optimum in an adaptive way (Sun et al., 2007).

2.2.2. Active learning

Active learning’s core idea refers to obtain cases that are difficult to mark out class to train one model. For active learning: firstly, the experts manually labelled the sample labels served as the initial training set, and then put it to use to learn the classifier. Secondly, some query algorithms were used to select samples that samples of one class are indistinguishable from other classes. And these samples are labeled by experts to expand the training dataset. Thirdly, the label samples are added to train a new classifier. After repeating step two and step three, qualified classifier is obtained. Merit of active learning is decreasing size of train samples, keeping main information, and reducing manual (Attenberg and Ertekin, 2013).

2.2.3. Decision adjustment learning

Decision Adjustment learning modifies the decision threshold, which is directly making positive compensation for the decision to correct the original unsatisfactory decision. In essence, it is an adjustment strategy, which makes the classification results tend to core range of the minority (Gao et al., 2020).

2.2.4. Feature extraction learning

Class imbalanced learning from feature selection driven refers to which the key features are preserved, which can increase the discrimination degree between the minority class and the majority classes, and improve the accuracy of the minority class and even any class. Feature extraction skills mainly include convolution neural network (CNN) and recurrent neural network (RNN) (Hua and Xiang, 2018). According to whether or not the evaluation criteria selected by feature selection are related to classifiers, three models have been developed: filter, wrapper and embedded (Bibi and Banu, 2015). These above ideas were noticed by researchers, and then series features of driven based algorithms were proposed (Shen et al., 2017; Xu et al., 2020). These algorithms have been applied to high dimensional data processing such as software defect (He et al., 2019), bioinformatics (Sunny et al., 2020), natural language processing (Wang et al., 2020) and network public opinion analysis (Luo and Wu, 2020).

2.2.5. Ensemble learning

Ensemble learning can first review the idea of cascade multi classification integration system written by Sebestyen. Ensemble learning is also one of the important technologies of machine learning. It can solve the limitations of some single algorithms by strategically building multiple base algorithms and combing them to complete classification task. One weak classifier that is slightly better than random conjecture can be promoted to a strong classifier by ensemble learning (Witten et al., 2017; Schapire, 1990).There are two leading frameworks for ensemble learning: one is Bagging framework (Breiman, 1996), and the representative algorithm is random forest algorithm (Verikas et al., 2011), and the other is Boosting framework (Ling and Wang, 2014; Li et al., 2013), and the representative algorithm is AdaBoost algorithm (Schapire, 2013).

Resampling-based ensemble learning，which is defined as an ingenious combination of resampling and ensemble learning. The simplicity of bagging paradigm firstly was noticed by researchers, and then multifarious algorithms have been developed, such as AsBagging algorithm (Tao et al., 2006) and UnderOverBagging algorithms (Wang and Yao, 2009). The former perfectly combines RUS with Bagging, and its merit is that it could reserve all cases of the majority class and reduce overfitting degree of the minority class. Meanwhile, AsBagging algorithm makes the classification result more stable because of the random resampling method and ensemble learning. Nevertheless, the result of the algorithm may is swinging with handing multi-noise datasets, and the reason is that the algorithm uses Bootstrap technology to create the train datasets of the basic algorithm. Thus, AsBagging_FSS algorithm (Yu and Ni, 2014) was proposed，which combined with the random feature subspace generation strategy (FSS). Because FSS can reduce the impact of noise samples on the basic classification algorithm, the classification results of the basic classification of the algorithm can get a better result. Therefore, AsBagging_FSS algorithm is better than AsBagging_FSS in dealing with the imbalanced data sets with noise samples. Except for combination of the resampling methods and Bagging ensemble learning framework, researchers also research the combination of the Booting framework and then develop some algorithms, such as SMOTEBoost algorithm (Chawla et al., 2003) and RUSBoost algorithm (Seiffert, 2010). Besides, the Hybrid framework (Galar, 2012) that was fusion of the Bagging and the Boosting was also noticed by researchers. Based on this idea, EasyEnsemble algorithm and BalanceCascade algorithm were proposed by Liu et al (Liu et al., 2009).EasyEnsemble algorithm is Bagging-based AdaBoost ensemble learning algorithm, which used Adaboost algorithm as basic classifier and first uses RUS algorithm to generate balanced train datasets of basic algorithm. EasyEnsemble algorithm can lower the variance and deviation of classification result, which makes the classification result become stable and presents stronger generalization ability. The BalanceCascade algorithm is improved EasyEnsemble algorithm. This algorithm's ingenious idea is that the correctly classified samples are constantly removed in the basic classifier train datasets, so that the classifier can repeatedly learn misclassified samples. Therefore, the generation of base classifier in the former algorithm is a parallel relationship, while the generation of base classifier in the latter algorithm is a serial relationship. Some representative algorithms are shown in Table 2.

Table 2. Representatives of ensemble learning methods.

	Algorithms	Ensemble frameworks	Combined strategies
Data driven	AsBagging	Bagging	Bootstrap
	UnderBagging	Bagging	Undersampling
	OverBagging	Bagging	Oversampling
	SMOTEBagging	Bagging	SMOTE
	SMOTEBoost	Boost	SMOTE
	RUSBoost	Boost	Undersampling
	EasyEnsemble	Hybrid	Undersampling
	BalanceCascade	Hybrid	Undersampling
Cost-sensitive	CS-SemiBagging	Bagging	Fuzzy cost matrix
	AdaCX	Boost	Empirical cost matrix
	AdaCost	Boost	Fuzzy cost matrix
	DE-CStacking	Stacking	Adaptive cost matrix

Ensemble algorithm is based on cost sensitive learning，which combines cost sensitive learning with ensemble learning. For example, AdaCX algorithm (Sun et al., 2007), which combines cost sensitive learning and AdaBoost algorithm, aiming to giving a larger weight to the minority class. The core of this algorithm is that update weights are different from different classes, which can be able to amplify effect of the cost sensitive, and AdaC1, AdaC2 and AdaC3 algorithm are developed based on different update weights. In addition, similar algorithms include AdaCost algorithm (Zhang, 1999), CBS1 algorithm and CBS2 algorithm (Ling, 2007). In addition, algorithms based on other frames are developed, such as the CS-SemiBagging algorithm of based Bagging ensemble framework (Ren et al., 2018), and the DE-CStacking algorithm (Gao et al., 2019) of based Stacking ensemble framework (Wolpert, 1992).

Ensemble learning algorithm is based on decision adjustment learning, among these algorithms, the classical algorithm is EnSVM-OTHR algorithm (Yu et al., 2015), which is the SVM-OTHR algorithm as the basic classifier and Bagging frameworks as the learning framework. EnSVM-OTHR algorithm uses bootstrap sampling and random interference to enhance the diversity of basic classifiers.

From the above analysis, we can draw a conclusion that ensemble learning can be applied to deal with the problem of class imbalance, especially for linear indivisible data, the ensemble learning presents better classification results. In the future, the ensemble-based class imbalanced learning methods will be one of the main research directions (Tsai and Liu, 2021). However, ensemble learning presents disadvantages of long training time and high computational complexity. Especially, it is also a bottleneck to deal with high dimensional and large scale data. Therefore, ensemble learning based on algorithms are facing with new challenges and opportunities in the era of big data. To solve this problem, ensemble learning can combine with feature extraction to reduce the data dimension. Or we deal with the problem of computational complexity by using the distributed Computing (Yang, 1997; Guo et al., 2018).

Table 3. Summary of data sets used in the experiment.

Data sets	Variates	Size	IR	Tra: Tes
yeast	8	514	9.08	7:3
glass	9	314	3.20	7:3
cleveland	13	177	16.62	7:3
vehicle	18	846	3.25	7:3

To sum up, the class imbalanced learning methods were analyzed from two different motivations. Although methods are from different ideas, the pursuit of the goal is consistent. Both data-driven methods and algorithm driven methods pursue the maximum accuracy of all classes. Therefore, methods from the data driven are essentially the same as the cost sensitive technology of some methods from algorithm driven. For example, in the random oversampling, in generating cases of the minority classes to balance quantity, it is also equivalent to giving the classifier a cost of IR times to the minority for some classifiers. The methods of manual resampling are similar to the idea of fuzzy cost sensitive algorithms, both of which use the prior information of samples to generate cases of the minority class or obtain the cost matrix.

Table 4. Results from DT classifier of oversampling-based and cost sensitive based.

Dataset	Item	Precision	Recall	F1-Score	Accuracy
yeast	**1**	0.9231	0.5714	0.7059	0.9355
	2	0.8462	0.6471	0.7333	0.9484
	3	0.9231	0.5714	0.7059	0.9355
	4	0.8462	0.5500	0.6667	0.9290
	average	0.8847	0.5850	0.7030	0.9371
	cost-sen	**0.92310**	**0.5714**	**0.7059**	**0.9355**
glass	1	0.8125	1.0000	0.8966	0.9538
	2	0.8750	1.0000	0.9333	0.8769
	3	0.9375	1.0000	0.9677	0.9846
	4	0.9375	1.0000	0.9677	0.9846
	average	0.8906	1.0000	0.9413	0.9500
	cost-sen	**0.8750**	1.0000	**0.9333**	**0.8769**
cleveland	1	0.6667	0.3333	0.4444	0.9038
	2	0.3333	0.2500	0.2857	0.9038
	3	0.3333	0.2000	0.2500	0.8846
	4	0.3333	0.2500	0.2857	0.9038
	average	0.4167	0.2583	0.3165	0.8990
	cost-sen	**0.3333**	**0.2500**	**0.2857**	**0.9038**
vehicle	1	0.8788	0.9062	0.8923	0.9449
	2	0.8333	0.9016	0.8661	0.9331
	3	0.8636	0.9194	0.8906	0.9449
	4	0.8485	0.918	0.8819	0.9409
	average	0.8561	0.9113	0.8827	0.94095
	cost-sen	**0.8182**	**0.9153**	**0.8640**	**0.93310**

Based on the above analysis, related experiments were designed:

- Experimental environment: Python 3.8.5 (64x), sklearn module, decision tree classifier (DT), default parameters.
- Dataset from keel website, shown in the Table 3, and the ratio of train set to test set is set to 7:3, which designates “Tra: Tes”.

We designed 10 oversampling experiments for that dataset, and randomly recorded the experimental results of four of them and calculated that average, numbered as “1”, “2”, “3”, “4”, and “average”. We also designed an empirical weighted cost sensitive experiment as a contrast, and the result of the experiment is numbered “cost-sen”. This conclusion has been obtained that cost sensitive experiment may catch similar classification results in oversampling experiments. Above analysis is shown in the Table 4.

After analyzing, we can acquire the general processing to imbalanced datasets, shown in Figure 2.

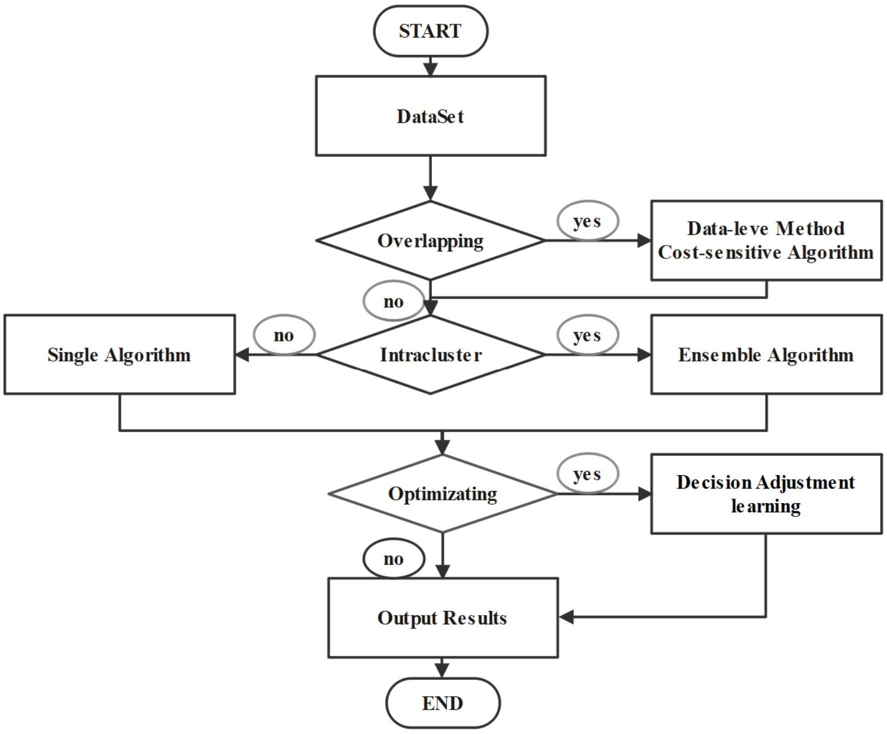


Figure 2. Flowchart of the handle to class imbalanced data.

Thus, we can draw the following conclusions. If one dataset is class imbalanced and non-overlapping, it is possible that standard classifiers are not affected. When it is overlapping in sample space to the dataset, it is difficult to categorize the sample of overlapping range; decision result is affected by the inverse probability theory which makes the decision results prefer the majority class. In this case, class imbalanced learning methods such as the data driven or algorithm driven method mentioned above can be employed. Or when the dataset has the phenomenon of interclass aggregation, it is difficult for a single classifier to distinguish the samples of the sub-aggregation range of the minority class. So, we can use ensemble-based class imbalanced learning to solve this data. In this way, it may improve the accuracy of all classes of the single classifier. In addition, when we get one classifier, we also can adjust the decision threshold by the decision adjustment learning according to experience, which may achieve better results. The whole process is illustrated in Figure 2.

3. Evaluation indexes

For result evaluation indexes of the different classifiers, a series of indexes such as threshold based, probability based and grade based can be found in some scientific literature (Luque et al., 2019). But some indexes of standard classifiers are unsuitable for the study file of the class imbalanced classifiers. Usually, we use robustness indexes such as F-measure, G-means metric, MCC and AUC. These based on confusion matrix indexes are creative.

Table 5. Confusion matrix of classification results.

	Prediction positive	Prediction negative
Positive class	True positive (TP)	False negatives (FN)
Negative class	False positive (FP)	True negatives (TN)

An explanation of the Table 5: TP (TN) is the number of samples that originally belong to the positive (negative) class and belong to the positive (negative) class after classification，which represents the number that is correctly classified; FP (FN) is the number of samples that originally belong to the negative (positive) class and belong to the positive (negative) class after classification，which represents the number of misclassifications.

A series of concepts such as Precision, Recall and TNP etc. are constructed by researchers:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$TNR = \frac{TN}{TN + FP} \tag{4}$$

$$G - Mean = \sqrt{TPR \times TNR} \tag{5}$$

$$F - Measure = Precision \times Recall \tag{6}$$

$$\mathrm{F1} - \mathrm{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)(TN + FN)(TN + FN)}} \tag{8}$$

Equation (3)’s “Recall” also can be called TPR.

G-mean is the geometric mean that is the accuracy of the positive class and the negative class. When the accuracy of the two classes is robustness, the G-Mean value becomes the optimal value. F-Measure is some similarity in the principle G-Mean. When the value of the Precision and Recall is roughly the same, the F-Measure value also becomes the optimal value. MCC represents the correlation degree between the real result and the predicted result, which is not affected by the class imbalance data. As one of correlation coefficients, the MCC value range is between −1 and 1. AUC is the area under ROC, and ROC is an important plane curve by FPR as the horizontal coordinate and TPR as the vertical coordinate.

4. Challenges and prospects

At present, class imbalanced learning methods have developed many mature methods in binary data, and a lot of algorithms and tools are used in various applications. In this era of big data, class imbalanced learning methods are facing some new challenges (Leevy et al., 2018; Chandresh et al., 2016):

- Large scale data processing problems: overcoming the increasing computational complexity and memory consumption.
- High dimensional data processing problems: sparse data processing.
- Data stream processing problems: the development of scalable online algorithms.
- Missing label data processing problems: semi supervised algorithm development.
- Multi class imbalance processing problems: the new definition of class imbalanced degree.
- Highly imbalanced processing problems: the development of accurate discriminant algorithms for the minority samples.

Nowadays, the processing of the class imbalanced problem is still research hotspot. The future research prospects are as follows:

- Strengthen theoretical research and enhance the interpretability of the algorithms. So far, there is a lack of theoretical research on class imbalanced model classification. It is difficult to interpret some the methods and evaluation is empirical.
- Adapt to the current research and be fit to the topical development. The complex data lead to the failure result of many traditional methods. Therefore, auxiliary technologies such as feature creation, feature extraction and active learning will be further applied in the study of the complex data.

5. Conclusions

In this research, we attempted to provide a review of methods in class imbalance problem. Different from other researches that have been published in imbalanced learning field, research are reviewed from both core technologies which are including the resampling methods and the cost sensitivity learning, and supporting technologies which include the active leaning and others. Through our analysis, we found some interesting conclusions flowingly:

- Data resampling based on classifiers are generally used in biomedical field due to the fact that biomedical data generally are fixed with structure and have multifarious similarity measurement between samples. Cost-sensitive learning technology is generally used in the operational research field, because its goal is to minimize the cost. With the improvement of data technology, data with high dimensionality and large scale are aroused by sensors. Feature extraction learning is used to reduce the complexity of some algorithms by reducing the dimension in high dimensional data. Distributed computing technology will be used to relieve the problem of insufficient memory in the single machine model in large scale data.
- The class imbalance rate is not an absolute condition that affects the result of the standard classifier. The standard classification model, in which the class of data is non-overlapping in the sample space, can also train outstanding result. When facing various datasets, researchers will choose the appropriate processing method according to the different data characteristics. For instance, when facing datasets with interclass aggregation factor, researchers often choose ensemble learning and complex classifiers that enable to distinguish examples of the secondary features of in

interclass. When facing datasets with fewer labels, researchers will choose semi-supervised, active learning and other supporting technologies to fit to the imbalanced dataset.

- The main challenge to fit to valid classifiers for class imbalanced datasets is the increasing complexity of data. For example, the processing of unstructured data such as language, text and web pages often needs data cleaning and feature representation. In addition, the handing of stream data generated by sensors requires developing dynamic learning algorithm with strong scalability and non-traditional memory.

At the end of this study, the future research directions are put forward from reviewing, which is also our focus in the future research.

Acknowledgements

This work was supported by NSSF of China (18BTJ029).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Attenberg J, Ertekin S (2013) Class Imbalance and Active Learning, In: He HB, Ma YQ, *Imbalanced Learning: Foundations, Algorithms, and Applications, IEEE,* 101–149.

Bibi KF, Banu MN (2015) Feature subset selection based on Filter technique. 2015 International Conference on Computing and Communications Technologies (ICCCT), 1–6.

Blagus R, Lusa L (2013) SMOTE for high-dimensional class-imbalanced data. *BMC Bioinf* 14: 1–6.

Breiman L (1996) Bagging Predictors. *Machine Learn* 24: 123–140.

Chandresh KM, Durga T, GopalanVV (2016) Online sparse class imbalance learning on big data. *Neurocomputing* 216: 250–260.

Chawla NV, Bowyer KW, Hall LO, et al. (2011) SMOTE: Synthetic Minority Over-sampling Technique. *J Artificial Intell Res* 16: 321–357.

Chawla NV, Lazarevic A, Hall LO, et al. (2003) SMOTEBoost: Improving Prediction of the Minority Class in Boosting. European Conference on Knowledge Discovery in Databases: Pkdd Springer, Berlin, Heidelberg, 20: 118–132.

Cmv A, Jie DB (2018) Accurate and efficient sequential ensemble learning for highly imbalanced multi-class data. *Neural Networks* 128: 268–278.

Dai HL (2015) Class imbalance learning via a fuzzy total margin based support vector machine. *Appl Soft Comput* 31: 172–184.

Domingos P, Pazzani M (1997) On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learn* 29: 103–130.

Galar M, Fernandez A, Barrenechea M, et al. (2012) A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE T Syst Man Cyb* 12: 463–484.

Gao HY, Lu HJ, Yan K, et al. (2019) Classification algorithm of gene expression data based on differential evolution and cost sensitive stacking ensemble. *Mini Comput Syst* 8: 66–78. (in Chinese)

Gao S, Dong W, Cheng K, et al. (2020) Adaptive Decision Threshold-Based Extreme Learning Machine for Classifying Imbalanced Multi-label Data. *Neural Process Lett* 3: 1–23.

Guo H, Li Y, Li Y, et al. (2018) BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. *Eng Appl Artificial Intell* 49: 176–193.

He H, Yang B, Garcia EA, et al. (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on IEEE, 1322–1328.

He H, Zhang X, Wang Q, et al. (2019) Ensemble Multi-Boost Based on RIPPER Classifier for Prediction of Imbalanced Software Defect Data. *IEEE Access*7: 110333–110343.

Hua Z, Xiang L (2018) Vehicle Feature Extraction and Application Based on Deep Convolution Neural Network. *Int J Eng Res* 7: 70–73.

Hui H, Wang WY, Mao BH (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Proceedings of the 2005 international conference on Advances in Intelligent Computing. Part I: 878–887.

Japkowicz N, Stephen S (2002) The Class Imbalance Problem: A Systematic Study. *Intell Data Anal* 6: 429–449.

Jing XY, Zhang X, Zhu X, et al. (2019) Multiset Feature Learning for Highly Imbalanced Data Classification. *IEEE T Pattern Anal* 9: 1–19.

Koziarski M, Woniak M, Krawczyk B (2020) Combined Cleaning and Resampling Algorithm for Multi-Class Imbalanced Data with Label Noise. *Knowl-Based Syst* 204: 1–17.

Krawczyk B, Koziarski M, Wozniak M (2020) Radial-Based Oversampling for Multiclass Imbalanced Data Classification. *IEEE T Neural Networks Learn Syst* 31: 2818–2831.

Kuang L, Yan H, Zhu Y, et al. (2019) Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor. *ITS J* 23: 161–174.

Leevy JL, Khoshgoftaar TM, Bauder RA, et al. (2018) A survey on addressing high-class imbalance in big data. *J Big Data* 1: 235–252.

Li K, Kong X, Zhi L, et al. (2013) Boosting weighted ELM for imbalanced learning. *Neurocomputing* 128: 15–21.

Li L, He H, Li J (2020) Entropy-based Sampling Approaches for Multi-Class Imbalanced Problems. *IEEE T Knowl Data Eng* 32: 2159–2170.

Li M, Xiong A, Wang L, et al. (2020) ACO Resampling: Enhancing the performance of oversampling methods for class imbalance classification. *Knowl-Based Syst* 19: 105–118.

Li YX, Yi C, Hu YQ, et al. (2019) Review of imbalanced data classification methods. *Control Decis* 34: 674–688. (in Chinese)

Lin J, Lu L (2021) Semantic Feature Learning via Dual Sequences for Defect Prediction. *IEEE Access* 9: 13112–13124.

Ling C (2007) A Comparative Study of Cost-Sensitive Classifiers. *Chinese J Comput* 7: 55–67.

Ling Y, Wang TJ (2014) Ensemble learning: a survey of boosting algorithms. *Pattern Recognit Artificial Intell* 01: 52–59.

Liu DX, Qiao SJ, Zhang YQ, et al. (2019) Survey of data sampling methods for imbalanced classification. *J Chongqing Univ Technol (NATURAL SCIENCE)* 033: 102–112. (in Chinese)

Liu XY, Wu J, Zhou ZH (2009) Exploratory Undersampling for Class-Imbalance Learning. *IEEE T Syst Man Cybern* 39: 539–550.

López V, Fernández A, García S, et al. (2015) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Info Sci* 250: 113–141.

Luo P, Wu B (2020) A big data dissemination feature mining system of Internet public opinion based on artificial intelligence. *Modern Electron Technol* 43: 184–187. (in Chinese)

Luque A, Carrasco A, Martín A, et al. (2019) The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit* 9: 216–231.

Maurya CK, Toshniwal D (2018) Large-Scale Distributed Sparse Class-Imbalance Learning. *Infor Sci* 456: 1–12.

Ogura H, Amano H, Kondo M (2011) Comparison of metrics for feature selection in imbalanced text classification. *Expert Syst Appl* 38: 4978–4989.

Ping R, Zhou SS, Li D (2020) Cost sensitive random forest classification algorithm for highly unbalanced data. *Pattern Recognit Artificial Intell* 33: 62–70. (in Chinese)

Pouyanfar S, Chen SC (2015) Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning. *Int J Semantic Comput* 11: 85–109.

Ren F, Cao P, Wan C, et al. (2018) Grading of diabetic retinopathy based on cost-sensitive semi-supervised ensemble learning. *J Comput Appl* 7: 2124–2129.

Rodriguez JA, Rui X, Chen CC, et al. (2013) Oversampling smoothness (OSS): an effective algorithm for phase retrieval of noisy diffraction intensities. *J Appl Crystallogr* 46: 312–318.

Schapire RE (1990) The Strength of Weak Learnability. *Machine Learn* 5: 197–227.

Schapire RE (2013) Explaining AdaBoost. *Empir Inference* 09: 37–52.

Seiffert C, Khoshgoftaar TM, Van J, et al. (2010) RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE T Syst Man Cyber* 40: 185–197.

Shen J, Xia J, Yong S, et al. (2017) Classification model for imbalanced traffic data based on secondary feature extraction. *IET Commun* 11: 1725–1731.

Sun Y, Kamel MS, Wong KS, et al. (2007) Cost-Sensitive Boosting for Classification of Imbalanced Data. *Pattern Recognit* 12: 3358–3378.

Sunny M, Afroze N, Hossain E (2020) EEG Band Separation Using Multilayer Perceptron for Efficient Feature Extraction and Perfect BCI Paradigm. 2020 Emerging Technology in Computing Communication and Electronics (ETCCE), 1–6.

Tao D, Tang X, Li X, et al. (2006) Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE T Pattern Analy Machine Intell* 7: 1088–1099.

Tao L, Huang YP, Wen Z, et al. (2019) The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) IEEE, 370–374.

Tsai CF, Lin WC (2021) Feature Selection and Ensemble Learning Techniques in One-Class Classifiers: An Empirical Study of Two-Class Imbalanced Datasets. *IEEE Access* 9: 13717–13726.

Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: A survey and results of new tests. *Pattern Recognit* 44: 330–349.

Wan JW, Yang M (2020) Review of cost sensitive learning methods. *Acta software Sinica* 31: 117–140. (in Chinese)

Wang D, Su J, Yu H (2020) Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language. *IEEE Access* 8: 46335–46345.

Wang S, Minku LL, Yao S (2015) Resampling-Based Ensemble Methods for Online Class Imbalance Learning. *IEEE T Knowl Data Eng* 27: 1356–1368.

Wang S, Yao X (2009) Diversity analysis on imbalanced data sets by using ensemble models. 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 324–331.

Wang T, Li ZJ, Yan YJ, et al. (2017) Survey of data stream mining classification technology. *Comput Res Dev*11: 1809–1815. (in Chinese)

Wang Z, Wu CH, Zheng KF, et al. (2019) SMOTETomek-Based Resampling for Personality Recognition. *IEEE Access* 8: 129678–129689.

Witten IH, Frank E, Hall MA, et al. (2017) Ensemble learning, In: Witten IH, Author, *Data Mining (Fourth Edition)*, 4 Eds., San Mateo: Morgan Kaufmann Press, 479–501.

Wolpert DH (1992) Stacked generalization. *Neural Networks* 2: 241–259.

Wu YX, Wang JL, Yang L, et al. (2019) A review of cost sensitive deep learning methods. *Comput Sci* 46: 8–19. (in Chinese)

Xiao LJ, Gao MR, Su XN (2019) An undersampling ensemble imbalanced data classification algorithm based on fuzzy c-means clustering. *Data Anal Knowl Discovery* 30: 90–96.

Xu Q, Lu S, Jia W, et al. (2020) Imbalanced fault diagnosis of rotating machinery via multi-domain feature extraction and cost-sensitive learning. *J Intell Manuf* 14: 1467–1481.

Yang Y (1997) A Comparative Study on Feature Selection in Text Categorization. Processing International Conference Machine Learning. 9: 73–85.

Ye ZF, Wen YM, Lu BL (2019) A review of imbalanced classification. *J Intell Syst* 4: 148–156.

Yu H, Mu C, Sun C, et al. (2015) Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data. *Knowl-Based Syst* 5: 67–78.

Yu H, Ni J (2014) An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data. *IEEE/ACM T Comput Biology Bioinf* 11: 657–666.

Yu H, Sun C, Yang X, et al. (2019) Fuzzy Support Vector Machine With Relative Density Information for Classifying Imbalanced Data. *IEEE T Fuzzy Syst* 27: 2353–2367.

Yu HL (2016) Basic idea and development of sample sampling technology, In: Yu HL, Author, Class imbalance learning theory and algorithm, 1 Eds., Beijing: Tsinghua University Press, 133–136.

Yu HL, Sun CY, Yang WK, et al. (2016) ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. *Knowl-Based Syst* 9: 55–70.

Zhai Y, Yang BR, Qu W (2010) Review of imbalanced data mining. *Comput Sci* 37: 27–32.

Zhang J (1999) AdaCost: Misclassification Cost-sensitive Boosting. Processing International Conference Machine Learning, 97–105.

Zhou ZH, Liu XY (2010) On Multi-Class Cost-Sensitive Learning. *Comput Intell* 26: 232–257.

Zong W, Huang GB, Chen Y (2013) Weighted extreme learning machine for imbalance learning. *Neurocomputing* 101: 229–242.