Necessary Information to Know to Solve Class Imbalance Problem: From a User's Perspective

G. Rekha and Amit Kumar Tyagi

Abstract In many real-world applications, class imbalance problem is the most attentive (also a major challenging) problem for machine learning (ML). The traditional classification algorithms assume evenly distributed in the underlying training set. In class imbalanced classification, the training set for one class called (majority class) far exceed the training set of the other class called (minority class), in which, the more often interesting class is minority class. We need to increase minority class samples than majority class samples in analysing of a datasets (related to an application). This is a hot problem in the past several decades. This article tries to provide as much as information to know or work about class imbalance problem with a detail description (from a user's perspective). For this, we include several articles from a reputed publication like IEEE, ACM, Elsevier, Wiley, etc. Hence, this work will help a lot to all the future researchers to find out or a summary (about their interest) with respect to this class imbalance problem (raising in several applications).

Keywords Class imbalance · Majority class · Minority class · Training set

1 Introduction

Recently, in machine-learning and data-mining domain [1], the class imbalance problem has been recognized a serious problem, which required solution with a higher accuracy, i.e. sampling of datasets of respective classes should be highly accurate. This problem occurs when the training data is not equally distributed among different classes. In several real-world's applications, such as like fault diagnosis, anomaly

G. Rekha (✉)
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Andhra Pradesh, 522502, India
e-mail: gillala.rekha@klh.edu.in

A. K. Tyagi
Department of Computer Science and Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana 121002, India
e-mail: amitkrtyagi025@gmail.com

P. K. Singh et al. (eds.), *Proceedings of ICRIC 2019*, Lecture Notes in Electrical Engineering 597, https://doi.org/10.1007/978-3-030-29407-6_46

detection, medical diagnosis, detection of oil spillage in satellite images, face recognition, text classification, protein sequence detection, and many others [2], the existence of this problem is crucial. The standard classifiers generally perform poorly in these cases. Classifiers usually ignore the minority class by overwhelming the majority class. But the minority class is often of more interest and more importance, therefore, it called for a strong urgency to be recognized. In the past decade, several works have been done with respect to class imbalance problem [3]. These methods can be categorized into three groups:

i. External approach: In the external approach, the data is pre-processed in order to balance the dataset to overcome the effect caused by imbalanced set.
ii. Internal approach: proposing new algorithms or change existing ones to handle the class imbalance problem at the time of classification.
iii. Hybrid approach: Combination of both external and internal approaches.

The drawback of internal approaches is of being algorithm explicit, whereas external approaches are independent of the classifier used. In general, class imbalance problem has been identified in several application areas/domains varying from education, biomedical, management, and many more. The solutions proposed in past are helpful but not enough to address this problem. The techniques proposed are highly depending on the nature of the data in a skewed distribution.

Hence, the remaining paper is organized as: Section. 2 summarizes existing analyzing techniques to solve the class imbalance problem. Further, several domains (applications) have been noticed (included) with respect to imbalance data (having class imbalance problem) in Sect. 3. Then, Sect. 4 discusses several evaluation models for addressing the class imbalance problem in brief. Further, Sect. 5 describes several possible research directions. Later, we discussed this problem (class imbalance problem) in Sect. 6 with one (at least) real-world example with respect to user's perspective and get several interesting facts or things (like accuracy paradox, overfitting, etc.). In last, this work is concluded with some future remarks in Sect. 6.

2 Existing Analysis Techniques for Solving the Class Imbalance Problem

In the past decade, many algorithms have been proposed by several researchers to address (or solve) the class imbalance problem in classification. In this section, we present an overview of all the techniques for imbalanced learning techniques. Broadly, the three basic strategies for dealing with class imbalance learning are preprocessing, algorithm-centric and ensemble methods. In the pre-processing approach, resampling of sample space is done to balance the data. Apart from resampling, the feature selection method is also adopted to optimize the feature space.

Pre-processing approach: Before building the classification model, the data is preprocessed to balance the imbalanced data. Under preprocessing, resampling of data is done to obtain the balance in the sample space [4].

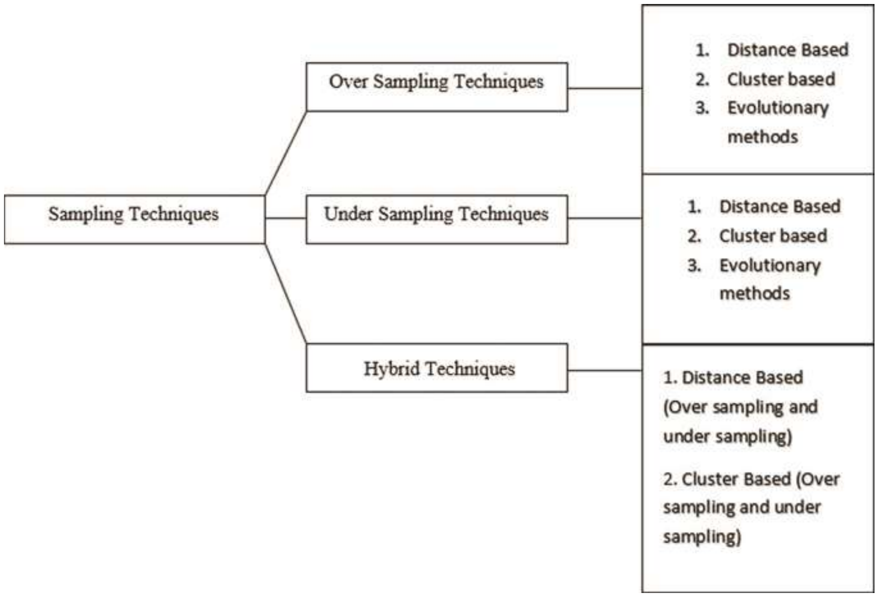

Fig. 1 Different sampling methods to solve the class imbalance problem

In general, the resampling techniques classified into three categories depending on the balancing techniques used to handle skew distribution. Figure 1 discusses respective three categories in brief.

- Under-sampling Technique: The imbalanced dataset is balanced by removing the irrelevant samples in the majority class. Random under-sampling (RUS) is the most effective (and mostly used in under-sampling) and simplest sampling technique [5]. Generally, RUS randomly eliminates the samples, a collection of datasets (i.e., which belongs to majority class), to balance the skew distribution in the respective datasets or with minority class's samples.
- Over-sampling Technique: An over-sampling technique generates synthetic data samples for the minority class to balance the skew distribution. Random over-sampling (ROS) and synthetic minority over-sampling technique (SMOTE) are simple and widely used techniques [6].
- Hybrid Technique: The combination of both over-sampling and under-sampling approaches is used to balance the skew distribution.

Feature Selection: **Compared to Pre-processing techniques**, significantly fewer work was done in past considering feature selection for class imbalance problem. In general, the aim of feature selection is to select a subset of *k* features from the entire feature space. The selection of important features will result in the optimal performance of the classifier [7]. Here, feature selection is divided into three processes, i.e., filters, wrappers. and embedded methods (see Fig. 2).

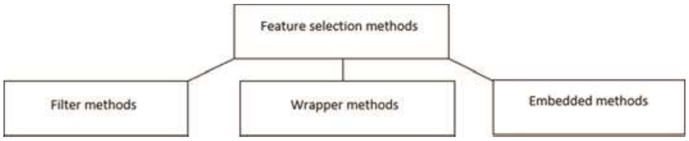

Fig. 2 Different feature selection techniques

Algorithm-Centric Approach: To improve the classification performance for imbalanced data, an attempt by the research community was to improve the learning ability of the existing classification algorithm. In the past decade, a plenty of improved classifiers such as support vector machine (SVM) [8], neural network (NN), *k*-nearest neighbor (KNN), decision tree (DT), rule-based classifiers, Naïve Bayes (NB), and many more have been proposed for handling imbalanced datasets in classification. Hence, all techniques/algorithms (as an interrelation) are discussed or showed in Fig. 3.

Ensemble Methods: Ensemble methods or classifiers, combine several base classifiers to improve the performance than using a single classifier. Nowadays, among all existing methods, one popular method is used mostly to solve this class imbalance problem, i.e., which is called as ensemble classifiers. The different ensemble techniques are bagging, boosting, and hybrid methods. While implementing ensemble methods [7], the different base classifiers used in the literature are support vector machine (SVM), neural network (NN), NB, decision trees including C4.5, CART, random forest, and rule-based classifiers.

Further from Fig. 4, we can say that boosting-based method includes techniques

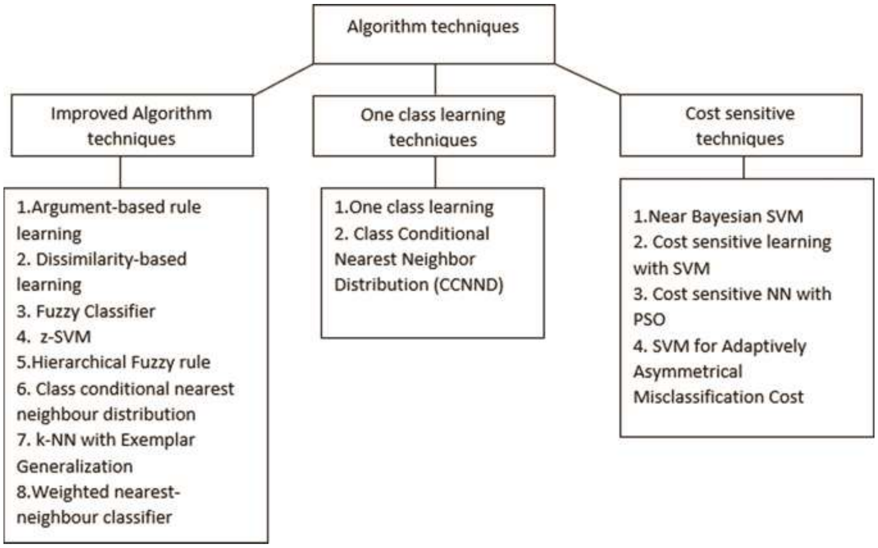
Algorithm techniques
Improved Algorithm techniques
One class learning techniques
Cost sensitive techniques
1.Argument-based rule learning
2. Dissimilarity-based learning
3. Fuzzy Classifier
4. z-SVM
5.Hierarchical Fuzzy rule
6. Class conditional nearest neighbour distribution
7. k-NN with Exemplar Generalization
8.Weighted nearest-neighbour classifier
1.One class learning
2. Class Conditional Nearest Neighbor Distribution (CCNND)
1.Near Bayesian SVM
2. Cost sensitive learning with SVM
3. Cost sensitive NN with PSO
4. SVM for Adaptively Asymmetrical Misclassification Cost

Fig. 3 Different algorithm-centric approaches

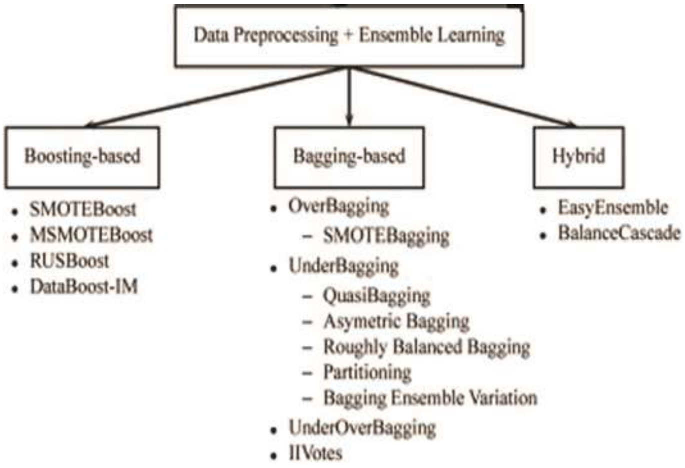

Fig. 4 Different ensemble methods

like SMOTEBoost, MSMOTEBoost, RUSBoost, and DataBoost-IM, whereas, bagging-based method includes several techniques like OverBagging, Underbagging, and UnderOverBagging, etc. In last, the hybrid method includes only two methods (till now) which are EasyEnsemple and BalanceCascade.

Hence, this section discusses several existing analysis techniques to solve the class imbalance problem. Now, in the next section, we will discuss about several domains (applications) with respect to class imbalance data.

3 The Existence of Imbalanced Data Classification in the Different Application Domains

The data-mining and machine-learning methods have been widely used in automated systems for analyses as a massive amount of data are generated currently. The prediction of the future actions based on past actions is an important problem. Prediction often consists of rare actions. Such rare actions data are usually an imbalance in nature. More efforts are required by the learning algorithms to learn such rare elements in the data. Class imbalance problem is existing for different application domain ranging from engineering to management (see Table 1).

Hence, this section discusses about several interesting applications (related to real world) which have imbalance problem of dataset (in a large dataset). Now, next section will discuss about several evaluation models for addressing class imbalance problem.

Table 1 Imbalanced data problem with respect to several applications

Domain	Application	Approach	Use cases
Biomedical engineering	Medicine and Biology for healthcare	Detection and prediction of abnormal structure in chemical and biomedical activities	Disease diagnoses, early warning of diseases, and detection of proteins [9]
Chemical engineering	Physical and life sciences	Conversion of cells into useful forms	Prediction of gene expression and drug resistance [10]
Financial management	Financial problems	Planning, procuring and control of resources of financial firms	Fraud detection in Credit card application [11], stock market prediction, and credit loan application
Information technology	Information devices	Identifying interesting events generated from information devices	Software defect prediction [12], evaluating the quality of software, and network intrusion detection
Web technology	Web application	Analyzing the Web data	Sentiment analysis, QoS prediction of Web services [13], detection of fake Websites
Energy management	Energy consumption and production	Planning and operation of energy consumption and production units	Availability of power distribution system [14], fitness levels of power transformers, fraud detection in energy field
Security	Security management system	Effective planning, organizing and controlling of security measures	Risk detection, analysis of risk in identifying crimes, and detecting unusual events from video surveillance [15]
Education	Education systems	Effective understanding of educational issues	Detection of student failure rate [16]

(continued)

Table 1 (continued)

Domain	Application	Approach	Use cases
Business	Business management system	Planning, organizing, and controlling an organization as while as staffing, leading	Customer relationship management (CRM) [17] data to analyse the customer behavior, preferences and need, detection/prediction of customer churn
Industry	Industrial manufacturing system	Identifying and solving the manufacturing problems pertaining to industries	Machinery fault diagnosis and defect detection in an automated inspection system [18]

4 Evaluation Model for Addressing Class Imbalance Problem

As disused above, class imbalance problem consists two classes majority and minority, whereas, we need to focus to increase balance among minority class (positive), i.e., among an imbalance dataset, we need strongly to balance minority class of data as an urgency to get accurate results. In general terms, it is a popular problem in artificial intelligence/machine learning, where the total number of samples related to one class (i.e. positive or minority class) is far less than the total number of another class (i.e. negative or majority class).

Samples of Minority Class < Samples of Majority Class

So, we used several techniques to solve this (class) imbalance problem using changing class distribution (using under-sampling techniques, over-sampling and advanced sampling), features selection, classifiers level (via manipulating classifiers internally, cost-sensitive learning, one-class learning), and ensemble method (refer Sect. 2). The standard metric to measure the classifier performance is by using confusion matrix (cost matrix) as shown in below table.

The acronym for each cell in the confusion matrix (represented in Table 2) is as following:

Table 2 Confusion matrix

	Predicted	
Actual	TP	FN
	FP	TP

Table 3 Metrics used in confusion metrics

Metrics	Formula
Accuracy	(TP + TN)/(TP + FN + TN + FP)
Error rate = 1-Accuracy	(FP + FN)/(TP + TN + FP + FN)
True positive rate (TPrate)/Recall/sensitivity	TP/(TP + FN)
True negative rate (TN rate)/specificity	TN/(TN + FP)
False positive rate (FP rate)	FP/(TN + FP)
False negative rate (FN rate)	FN/(TP + FN)
Precision/positive predictive value (PPvalue)	TP/(TP + FP)

- TP (True Positive): The number of positive samples correctly classified as positive.
- TN (True Negative): The number of negative samples correctly classified as negative.
- FP (False Positive): The number of negative samples incorrectly classified as positive.
- FN (False Negative): The number of positive samples incorrectly classified as negative.

These terms (i.e. metrics TP, TN, FP, and FN) have been discussed in Table 3. These metrics are the mostly used metrics for evaluating the performance of the classifiers. Among them, accuracy and error rate are mostly used one-performance metric among all metrics (i.e. to solve class imbalance problem). Accuracy assesses the overall efficiency of the classifier for binary class problems.

- Precision/positive predictive value is used to measure the correctness of the positive samples (out of total positive samples how many are correctly classified as positive).
- True Positive Rate (TP rate)/Recall/Sensitivity is a measure of completeness or it specifies the accuracy of positive samples.
- True Negative Rate (TN rate)/Specificity is a measure of actual negative samples correctly identified as negative.
- False Positive Rate (FP rate) is measured as number of incorrect positive predicted samples divided by the total number of negative samples. It is calculated as 1-specificity.
- False Negative Rate (FN rate) is the measure of proportions of positive samples, which are incorrectly classified as negative samples.

Considering the class-imbalanced datasets, accuracy gives more importance to the majority/negative classes than the minority classes. This leads to a difficulty for classifier to learn from minority class. For example, if we consider a problem where only 1% of the samples belong to the positive class (or minority class), high accuracy of 99% is achievable by predicting the negative class (or majority class) for all samples. Yet, all minority class samples, the rare, and more interesting samples, are misclassified. Moreover, , in the past decade, additional metrics were proposed for imbalanced datasets.

F-measure: It is also known as *F* Score/*F*1 Score, and also it can be defined as "a weighted harmonic mean of precision and recall." It can be defined as:

$$F\text{−measure} \ = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} \ * \text{Precision}}$$

The value of *F*-measure increases proportionally with an increase of precision and recall. The higher the value of *F*-measure, the better the classifier working on positive samples.

Geometric Mean (G-Mean): *G*-Mean is an interesting measure because it computes the prediction of the accuracies of the two classes, attempting to maximize them with good balance. It takes into consideration the sensitivity and specificity and provides a balance between classification performance on both majority and minority classes.

$$G\text{−mean} = \sqrt{\text{Sensitivity} * \text{Specificity}}$$

The prediction of poor performance on the positive samples may lead to low value of *G*-mean in spite of negative samples are correctly classified by the classifier. Two popular graphical tools used in imbalanced domains are the receiver operating characteristics (ROC) curve and the corresponding area under the ROC curve (AUC). Note that Provost et al. [19] proposed ROC and AUC as alternatives to accuracy.

Receiver Operating Characteristics (ROC) curve: The ROC curve gives graphical trade-off between TPrate (benefit) and FPrate (the cost). The curve gives true positive rate as a function of false positive rate. The performance of a classifier is represented by a single point in the ROC space. The more inclined the curve is toward the upper left corner, the better is the classifier's ability in discriminating positive (minority) and negative (majority) classes. The ROC curves are good ways to compare models or set of models.

Area Under Curve (AUC): The area under the ROC curve is a summary indicator of ROC curve. It provides the evaluation of the best model by sorting the performance of all models, on average. Higher the value of AUC, better the prediction model. The below Table 4 shows the scale for the interpretation of AUC.

Hence, this section discusses several evaluation models in Table 4 and metrics like ROC, *G*-Mean, etc., to measure performance of mechanism to solve class imbalance problem. Now, next section will investigate several future directions with respect to class imbalance problem.

Table 4 Scale for the interpretation of AUC

AUC value	Model performance
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Fair
0.5–0.6	Poor

5 Possible Research Directions

Today, solving class imbalance problem in a feasible and valid way is a challenging task. A reach problem is needed to solve in an affordable and accurate manner. Also, a research problem is needed to help someone (society) with its solution. Then we call that particular problem is a feasible and correct problem. Keeping this thing in our mind, in this section, we explain several possible research directions in class imbalance problem.

- **Ensemble Learning**: Ensemble-based algorithms have been employed to improve the class imbalance problem. The performance of the ensemble learning algorithms is based on the individual classifiers accuracies, and also the diversity in handling the single problem between all classifiers. The diversity is measured as a degree to which the classifiers make different decisions on one problem. In [20], authors considered diversity while building ensemble classifiers. But careful study still needed to address the diversity problem in the future.
- **Variation in Imbalanced datasets**: in the past decade, hundreds of algorithms proposed to handle the class imbalance problems and they demonstrated outstanding performance. All the algorithms in general treated imbalance datasets constantly and handled it using versatile algorithms. But, in real-world scenario, the imbalanced datasets have variation in terms of imbalanced ratio, a number of dimension/features, and the number of classes. In [7], the author suggested that using a specific learning algorithm is inefficient to handle such variations in the data. So, adaptive learning is the new research direction in class imbalance problem.
- **Imbalance Data Streams**: The availability of massive amount of data leads to big data analytics. The most important challenge is to handle and answer to streaming and fast-moving data. Online learning has received higher attention from community/researchers belonging to machine-learning and data-mining domain/area. So, when learning from an online stream, the main difficulty arises is online class imbalance learning. In [21] the author discussed the problems arise with online class imbalance problems like lack of prior knowledge in identifying minority data classes and majority classes. Since the requirements of quick and accurate responses needed for data that arrive at a different time interval in big data era, imbalanced data stream becomes a new and popular research direction.

- **Semi-Supervised Learning**: In most of the data analysis, collection of massive data is cheaper but it is expensive to label the samples to train the classifier. In big data, it is quite common to find massive data with few labeled samples and plentiful unlabeled samples. For class-imbalanced datasets, it is difficult and expensive for semi-supervised learning to extract class label for minority samples. Active learning is used in general to predict the label of unlabeled samples. For handling imbalanced datasets, few active learning algorithms have been proposed in past. More research work needed in investigation of different active learning for selecting and utilization of informative samples in imbalanced datasets.
- **Application Areas**: The application in which the imbalanced datasets exist are management science, decision-making, emergency management, natural disasters (public health incidents/social security incidents), IoT (Internet of things), Big Data, and security management.

Hence, this section discusses about several research directions related to class imbalance problem. Now, next section will discuss a real-world problem and will share several interesting facts (regarding to class imbalance problem) form a user's perspective.

6 Open Discussion

The class imbalance problem is typically a classification problem with the skewed distribution of class labels. It can be a binary classification problem (with two class labels) or multi-class classification problem (with more than two class labels). As discussed in Sect. 4, smaller class (minority or positive) is highly interested to balance to get accurate results. But the question is how we will identify which class is majority or minority. So, for that, this section provides a real-world example to under this problem in clear-cut manner. For example, in a medical diagnosis of a rare disease where there is a critical need (or huge requirement) to identify such a rare medical condition among the normal populations. Any errors in diagnostic will bring stress and further complications to the patients. The physicians could not afford any incorrect diagnosis since this could severely affect the patients' well-being and even change the course of available treatments and medications. Thus, it is crucial that a classification model should be able to achieve higher identification (i.e. accuracy) rate on the rare probabilities of occurring diseases/occurrences (minority class) in datasets. In another interesting example, suppose a patient visits to a doctor and he takes several features in consideration and based on that doctor told that you will die tomorrow. Then this patient visits to a different doctor and this doctor uses similar features (which first doctor used) then he concluded that you are ok. Now patient is confused. Now, if this patient visits to a third doctor and this doctor also used similar features then he may get a different result and will tell to patient that you may or may not die tomorrow (i.e., 50–50 chances). So, for such false positive, true positive, false negative, true negative, etc., we require confusion metric (see Sect. 4) to represent such decisions/data.

In another example, suppose a new disease is raised somewhere, where about 0.01% of people have contracted the disease. Then, we collect data of 60,000 people, which includes six people with the disease. Now here, if we want to make a model that, given data of a new person, predicts whether that person has contracted the disease, we can easily build a model that scores really high for accuracy on the training set. Our model easily can show/predict that no person has the disease or no person who is not exactly equal to one of our known disease carriers. But based on just predicting that everyone is healthy has an accuracy of 99.99%. Then such models will become useless (due to less sampling of data). Apart from that, if we re-sample and analyze this data again then this time we will get different accuracy results. Hence, accuracy is a paradox for users, i.e. getting high accuracy does not mean that classes are balanced now. So, we need to make our models like that they will be able to train and find patterns in our imbalanced data. So, in the future/whenever we have imbalanced data/classes, our model can learn to choose the majority class instead of reacting to the data. Also, we can try several other processes (can avoid accuracy paradox) like collecting more data for sampling, changing several metrics for respective data, or resampling our data again and again (several times, we did not get two consecutive similar results), generating synthetic samples, using different–different algorithms to balance our data, etc. Moreover this, there are several approaches (to solve class imbalance problems) like balancing class imbalance (via training a set with equal numbers of classes by sampling out surplus class data points), anomaly detection (taking smaller class as anomalies and using such algorithms for classifications), and generating minority class samples (using the distributions of features in minority class, sample out more datapoints). Note that some algorithms are less prone to the problems typically associated with unbalanced classes, as long as there are some samples of each class in the training data. As discussed above, accuracy is a paradox in machine-learning or data-mining domain, here it means "when our model's accuracy measures and give a response to us that we have received excellent accuracy (such as 90%), but in actual, this received accuracy is only reflecting the underlying class distribution." With a large number of dataset and analysing this datasets with more number of features, we may get good accuracy than the less number of datasets with consisting similar (or less) features. In this, classification accuracy is the first metrics, we use to solve our classification *problems*. Note that a false sense of performance can create overfitting in a trained data.

Hence, this section shows a clear-cut vision to class imbalance problem and effect of good accuracy (with less resampling of data) and overfitting in datasets. Now, next section with conclude this work in brief and will show a future path to the future researchers (interested one) in this respective (hot) problem.

7 Conclusions

As discussed above, we reached to a conclusion that class imbalance problem has identified as biggest problem (especially in machine learning) to solve in many practical applications/domains. Today's machine learning is used to learn several facts/extracting knowledge from imbalanced data sets. In Sect. 2, we have discussed several techniques to balance/solve this (class) imbalance problem using changing class distribution (using under-sampling techniques, over-sampling and advanced sampling), features selection, classifiers level (via manipulating classifiers internally, cost-sensitive learning, one-class learning), and ensemble method (refer Sect. 2). Hence, the purpose of this paper is to show a clear understanding to a reader/researcher to find related facts regarding/required to solve a class imbalance problem. This work also explains several domains where class imbalance problem arises and metric used to measure performance of techniques (which is used to solve a respective problem). Also, we end this work by showing several future research directions to the future researchers (with respect to respective problem). Remember that, sometime we may not get much improvement or much accurate result (performance of the induced classifier) through balancing an imbalanced dataset. Also, many learning systems are not much interested to find differences in class distributions. Hence, we need a clearer and deeper understanding of how class distribution affects each phase of the learning process for more learners. A deeper understanding of the basics will help us to design better methods for dealing with the problem of learning with skewed class distributions.

Acknowledgements This research is funded by the Koneru Lakshmaiah Education Foundation, Lingaya's Vidyapeeth and Anumit Academy's Research and Innovation Network (AARIN), India. The authors would like to thank Koneru Lakshmaiah Education Foundation, Lingaya's Vidyapeeth and AARIN, India, an education foundation body and a research network for supporting the project through its financial assistance.

References

1. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling imbalanced datasets: a review. GESTS Int. Trans. Comput. Sci. Eng. **30**(No 1), 25–36 (2006)
2. Longadge, R., Dongre, S.: Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707 (2013)
3. Ali, A., Shamsuddin, S.M., Ralescu, A.L.: Classification with class imbalance problem: a review. Int J Adv Soft Comput Appl **7**(3), 176–204 (2015)
4. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**, 113–141 (2013)
5. Tahir, M.A., Kittler, J., Mikolajczyk, K., Yan, F.: A multiple expert approach to the class imbalance problem using inverse random under sampling. In: International Workshop on Multiple Classifier Systems, pp. 82–91. Springer, Berlin, Heidelberg (2009)

6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority oversampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Yijing, L., Haixiang, G., Xiao, L., Yanan, L., Jinling, L.: Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowl.-Based Syst. **94**, 88–104 (2016)
8. Tang, Y., Zhang, Y.Q., Chawla, N.V., Krasser, S.: SVMs modeling for highly imbalanced classification. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **39**(1), 281–288 (2009)
9. Herndon, N., Caragea, D.: A study of domain adaptation classifiers derived from logistic regression for the task of splice site prediction. IEEE Trans. Nanobiosci. **15**(2), 75–83 (2016)
10. Raposo, L.M., Arruda, M.B., de Brindeiro, R.M., Nobre, F.F.: Lopinavir resistance classification with imbalanced data using probabilistic neural networks. J. Med. Syst. **40**(3), 69 (2016)
11. Zakaryazad, A., Duman, E.: A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. Neurocomputing **175**, 121–131 (2016)
12. Tan, M., Tan, L., Dara, S., Mayeux, C.: Online defect prediction for imbalanced data. In: 37th IEEE International Conference on Software Engineering (ICSE), 2015 IEEE/ACM, vol. 2, pp. 99–108. IEEE (2015)
13. Li, Z., Bin, Z., Ying, L., Yan, G., Zhi-Liang, Z.: A Web service QoS prediction approach based on collaborative filtering. In: Services Computing Conference (APSCC), IEEE Asia-Pacific, pp. 725–731. IEEE (2010)
14. Xu, L., Chow, M.Y., Taylor, L.S.: Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification $ E $-algorithm. IEEE Trans. Power Syst. **22**(1), 164–171 (2007)
15. Wang, Y., Li, X., Ding, X.: Probabilistic framework of visual anomaly detection for unbalanced data. Neurocomputing **201**, 12–18 (2016)
16. Márquez-Vera, C., Cano, A., Romero, C., Ventura, S.: Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Appl. Intell. **38**(3), 315–330 (2013)
17. Lessmann, S., Voß, S.: A reference model for customer-centric data mining with support vector machines. Eur. J. Oper. Res. **199**(2), 520–530 (2009)
18. Cateni, S., Colla, V., Vannucci, M.: A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing **135**, 32–41 (2014)
19. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
20. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: IEEE Symposium on Computational Intelligence and Data Mining, CIDM'09, pp. 324–331. IEEE (2009)
21. Ghazikhani, A., Monsefi, R., Yazdi, H.S.: Online neural network model for non-stationary and imbalanced data stream classification. Int. J. Mach. Learn. Cybernet. **5**(1), 51–62 (2014)