A novel approach to handle class imbalance : A Survey

[1]Ms. Monica. Ochani, [2]Dr.S.D. Sawarkar, [3]Mrs. Swati Narwane
[1]Research Assistant, [2]Principal, [3]Assistant Professor
Datta Meghe College Of Engineering

Abstract - Machine learning is study of algorithms that a system uses to effectively perform a specific task. It depends on the patterns and inference instead of any instructions. In machine learning, majorly there is some level of class imbalance issue in real-world classification. This problem arises when each class does not make up an equal division of a data-set. It is essential to properly alter the metrics and methods to balance the data set goals. This means that many learning algorithms of machine learning have low predictive accuracy for the not often occurring class. In this paper, we shall discuss this problem and look in to different approaches used to solve the class imbalanced issue. This paper discusses the survey of different approaches done to improve the class imbalance issue in the data sets by learning about the data level approaches and the algorithm approaches. We have discussed the oversampling and undersampling methods to overcome the data imbalance problem.

Keywords - Class imbalance, data mining, machine learning, imbalance data, applications, classification, approach, algorithm, sampling.

I. INTRODUCTION

In machine learning, data mining is one of the most important branch. As the real world is getting exposed with new technologies the data is also increasing with increase in number of problems. These problems can be marked as volume and velocity of data. In imbalanced classes accuracy is not always true which is very standard problem classification in machine learning. There is always a difference in datasets with asymmetric ratio of observations in a class. Few examples of applications which have imbalanced data sets are: reports of medical diagnosis, finance industry etc,. The datasets faces imbalanced class distribution when one of the class is not sufficiently represented. Basically it means that the number of examples of the class which are less than half of the whole dataset is considerably smaller than the number of examples of the class which is sufficiently represented. [1]

In recent years the studies have grown emphasis on class imbalance. The classification for class There are many industries which are affected by class imbalance distribution Reported works in classifications for class imbalance distribution come in many ranges of domain applications like diagnosis of faults, abnormality detection, medical diagnosis, oil spill detection in images taken by satellites, face recognition, text classification, and many others. The most important challenges of the class imbalance issue is of pattern recognition and data mining. Same problem is seen in practical applications. [3]

If an individual has learned about machine learning and data science, he/she definitely understands that imbalanced class distribution is always seen in machine learning. This situation occurs only when the observations listed in a class have is very low than that present in the other classes. This problem is primarily in applications where detection of inconsistency is crucial like electricity thievery, unauthorised transactions in banks, identification of uncommon diseases, etc. In these scenario, the predictive model developed using machine learning algorithms could be incorrect. This is mainly due to Machine Learning algorithms which are mostly made to improve correctness by minimising the error. Hence, they do not take into account the class distribution or balance of classes. [15]

II. Challenges in imbalanced datasets

We have already mentioned above, that electricity thievery, unauthorised transactions in banks are affected by class imbalance problem. Electricity thievery is one of the largest form of theft in the world. The utility companies are getting inclined towards advanced analytics and machine learning algorithms so that they can identify the consumption patterns by which we can indicate the theft. Although, one of the biggest hurdle is the humongous data and its distribution. unauthorised transactions in banks are mostly lower than normal bank transactions i.e. accounting the data to approximately around 1-2 % of the total number of observations. The requirement here is to improve identification of the infrequent minority class so as to achieve the overall accuracy of the majority class. [1]

Machine Learning algorithms tend to generate deplorable classifiers when it comes across the imbalanced datasets. For any imbalanced data set, if we have to estimate the occurring of the event and if that belongs to the minority class and the rate is less than 5%, it is usually referred to as a rare event.[9]

As there is quick development of information technologies, industries have to face new challenges to inspect huge amounts of information of data. This data can be classified as volume of the data, velocity of the data, variety of data, veracity, valence and value of the data. Volume of the data is created and is increasing every second, due to a digitized world with lots of information data. Due to this a variety of data in different forms like text, audio-video files, images geographic data etc is received. As mentioned the data is created every second hence we have the velocity factor. It is the speed at which the data is

created and rate at which it travels from one point to other. These are the three major aspects that identify and describe the challenges in a dataset. Veracity names to the noise and abnormality in data. It is often the unmeasurable uncertainties and trustworthiness of data. Valence refers to the connectedness of data in the form of graphs, just like atoms. [10]
The standard machine learning techniques analysis methods give inaccurately performance when faces a imbalanced datasets. The standard classifier algorithms like Decision Tree and Logistic Regression have a inclination towards classes which have more number of observations. They tend to only estimate the majority class data. The characteristics of the minority class are treated as noise and are most of the time ignored. Thus, there is a high chances of misclassification of the minority class as compared to the majority class. [10]

III. Approaches to overcome class imbalance

There are various approaches by which we can solve the class imbalance problem. For this we should be aware of supervised classification. The objective of classification is to estimate explicit labels where input and output are known. Many approaches are developed to overcome the challenges faced in imbalanced data.

3.1. Data level approaches

Data-level approach or rather than an external techniques which engages the pre-processing step to rebalance the class distribution. These approaches have been developed which can be implemented during the pre-processing level. The most commonly used approach is resampling. This approach includes two techniques namely undersampling and oversampling of the dataset. Resampling technique can be done with or without the replacement of the dataset. If a dataset is balanced by removing the instances from majority class is known as undersampling. When we add similar instances of minority class to balance the class ratio then oversampling is achieved. We can do resampling with or without the replacement. We shall see each approach in details below. [1]

3.1.1. Resampling

The process of recreating the sample data from the actual data sets is called resampling. It can be either done by non-statistical estimation or by statistical estimation. In non-statistical estimation, we randomly draw samples from the actual dataset hoping that the data is divided in a similar division to the actual dataset. However, in statistical estimation, we estimate the parameters of the actual dataset and then drawing the subsamples. Hence, we can extract data samples that carry most of the information from the actual population. Thus the resampling technique statical as well as non statistical help us in drawing the samples when the data is imbalanced.[11][12]

3.1.2. Undersampling

Undersampling is a technique in which we randomly select the samples from the majority class and discard the remaining. We assume that any random sample approximately reflect the division of the data. In this method the goal is to balance the distributions in the class through a random elimination of majority class observations. This results in discarding the useful data that could be important for classifiers.
The k-nearest neighbor based approach is one of the frequent used approaches. In these approaches the sample set is selected and then is searched exhaustively in the entire dataset and it will select the k-NN and discard the other data. It is assumed that k-NN carries all the information that we need regarding those classes in this method.
Many other undersampling techniques are also available which are based on two different types of noise model hypotheses. In this technique we assume that the samples near the boundary are noise. Hence the noise will be discarded in order to obtain the maximum accuracy.
For another noise model, the assumption is that if the location of the majority class samples and the minority class samples are same then they are noise. If we discard these samples from the data then it creates a clear boundary that can help in classification .[10][11]

3.1.3. Oversampling

By using the undersampling method we achieve an equal distribution by eliminating majority class samples. In oversampling we do this by replicating the minority samples so that the distribution is equal and balanced. One more very common approach used in oversampling is SMOTE. This method helps to overcome the shortcomings of oversampling. It creates the new samples by introducing based on the distances between the point and its nearest neighbors. SMOTE also calculates the distances for the minority samples which are near the decision boundary and generates the new samples. This will affect the decision boundary to move away from the majority classes and remove the overfitting issue.[10][11]

32. Algorithm level approach

Generally, the algorithm-level approach can be labelled as dedicated algorithms which directly takes the imbalance distribution from the classes in the datasets. They are recognition-based one class learning classifications, cost-sensitive learning and ensemble techniques. We shall discuss each of them in the following subsections.[1][3]

3.2.1. Improved algorithm

This is a classification algorithm developed by researchers to manage the classification of datasets to handle the class imbalance issue. It is modified to fit the requirement to directly understand from the imbalanced class distribution. These type of algorithms understand about the distribution of the classes before extracting major information in order to develop a model based upon the target objective. [1][3]

There is also a research done on Fuzzy to address the classification of imbalanced datasets. Hierarchical Fuzzy rule uses a linguistic rule generation technique to build the base rule. We can extract the hierarchical rule base (HRB) is extracted from the rule base. After this the best collective rules from HRB are selected using algorithm. The one other proposed study on Fuzzy Classifier which uses frequency distribution to generate membership degrees to each class before they build the corresponding fuzzy sets. This gives a new different approach to conventional Fuzzy. As it is purely data driven while it depends on trial and error method in building the if-then rules.[1][3]

3.2.2. One-class learning

One-class learning algorithms are also called as recognition based methods. It works by the classifier on the characterization of the minority class. It understands only from the examples of minority class rather than trying to identify the different patterns from examples of majority class and minority class. However, an effective boundary threshold is the key point with this approach as a strict threshold will separate apart the positive examples (minority class) while a moderate one will cover some negative examples (majority class) in the decision boundary.[1]

3.2.3. Cost sensitive learning

The motivation for cost-sensitive learning is the distinct characters of domain applications with class imbalance datasets and misclassification cost being considered as equal by numerous traditional learning algorithm. The idea of cost-sensitive learning approaches is that an expensive cost is imposed on a classifier when a misclassification happens. For example a classifier assigns larger cost to false negatives compared to false positives thus emphasizing any correct classification or misclassification regarding the positive class.[12]

3.2.4. Ensemble method

Another option is Ensemble learning for class imbalance problem. In this method several classifiers are trained on training data and their analysis are approximated to give the final classification decision. Ensemble methods can be defined as boosting in general. In this approach bagging stands for Bootstrap Aggregation to decrease the prediction difference by generating more examples for training set from original data. [14]

Each k variations of the training set will have a k number of classifiers. This is done by a classifier which is used for each of these training set examples by a chosen machine learning algorithm. By combining the output all the classifiers, we get the results. Boosting methods carry out experiments on training sets using many different models to induce classifiers to give output. For wrongly classified examples, higher weights are assigned to each classifier. By using weighted average approach the outputs are then updated. The final decision is obtained by collaborating data from all classifiers.

3.2.5. Hybrid Approach

In recent years new technique of classification algorithms have been devised for handling class imbalance datasets. This approach is besides the one-class learning, cost-sensitive methods and ensemble approaches. More than one machine learning algorithms is employed to improve the classification quality, frequently by the hybridization along with other learning algorithms to get better results. The hybridization is developed with the idea to reduce the problem in sampling, feature subset selection, cost matrix optimization and polish the learning algorithms. The hybridize classifiers are used in order to improve classification qualities with class imbalance problem. [15]

IV. Advantages and disadvantages of the approaches

In the above section we have seen many approaches which can handle imbalance data issue. These approaches when implemented have their own pros and cons. Let us discuss in details in the below sub sections

4.1. Advantages

Classification performance is made better by using the data generation and boosting methods. The use of various different classifiers in ensemble method preserves the regularity with the training data which is a significant factor to ensure correctness. It helps in improving the execution time of the model and solve the memory problem by decreasing the number of samples when the training data set is more. There is no loss of information in oversampling. In order to adjust the division of sample data boosting is used to handle uneven division in datasets by assigning weight to examples. [9]

4.2. Disadvantages

There is a loss of information in undersampling technique. Oversampling rises the possibility of overfitting, as it makes same copy of the minority samples than sampling from the distribution of minority samples. Another problem encountered in oversampling method is that as the number of samples expands, the model becomes more complex of and this expands the running time of the models. One class algorithm is restricted only to certain learning algorithms. Cost-sensitive learning we do not know the real cost in many applications even if the dataset is balanced. It also has an issue of over-fitting during training. It is same as over-sampling technique and there is no difference the performance of both the approaches .[9]

V. Conclusion

In this paper, we have discussed the class imbalance problem and look in to different approaches used to solve it. We have also explored many different approaches to improve the class imbalance in the data sets, this includes learning about the data level approaches and the algorithm approaches. *This paper concludes that the oversampling and undersampling methods can*

be used for tackling the imbalance class problems. It is very important to balance the imbalance data with effective techniques and at the same time, cost factor should be given attention. The correct classifier techniques and performance evaluation metrics must be applied to achieve good results.

VI. References

[1] Aida Ali , Siti Mariyam Shamsuddi, and Anca L. Ralescu, "Classification with class imbalance problem:A Review", Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, November 2015

[2] Kwabena Ebo Bennin, Passakorn Phannachitta, Akito Monden, and Solomon Mensah "MAHAKIL:Diversity based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 2017

[3] Lina Gong , Shujuan Jiang , Lili Bo, Li Jiang, and Junyan Qian, "A Novel Class-Imbalance Learning Approach for Both Within-Project and Cross-Project Defect Prediction", IEEE TRANSACTIONS ON RELIABILITY, 2019

[4] Joffrey L. Leevy 1 , Taghi M. Khoshgoftaar 1 , Richard A. Bauder 1* and Naeem Seliya, "A survey on addressing high-class imbalance in big data", 2018

[5] Omer Sagi , Lior Rokach, "Ensemble learning: A survey", 2017.

[6] Cuicui Luo, "A comparison analysis for credit scoring using bagging ensembles", 2018

[7] Bing Zhu Bart Baesens Sichuan, "Benchmarking sampling techniques for imbalance learning in churn prediction", Article in Journal of the Operational Research Society · March 2017

[8] Nathalie Japkowicz and Shaju Stephen, "The class imbalance problem: A systematic study" , 2018

[9] HUALONG YU , CHANGYIN SUN QI WANG , AND XIAOYAN XI , "A Fast and Flexible Cost-Sensitive Learning Framework for Classifying Imbalanced Data", June 19, 2018.

[10] Neelam Rout, "Handling Imbalanced Data: A Survey", January 2018

[11] Krystyna Napierala, Jerzy Stefanowski , "Types of minority class examples and their influence on learning classifiers from imbalanced data", 2015

[12] Mateusz Buda, Atsuto Maki, Maciej A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks", 2018

[13] Shuo Wang , Leandro L. Minku, and Xin Yao, "A Systematic Study of Online Class Imbalance Learning With Concept Drift", 2018

[14] Hartono, Opim Salim Sitompul , Tulus, , Erna Budhiarti Nababan, "Biased support vector machine and weighted-SMOTE in handling class imbalance problem", 2018

[15] Stjepan Picek, Annelie Heuser, Alan Jovic, Shivam Bhasin, Francesco Regazzoni, "The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-channel Evaluations", 2018

[16] Zhenbing Liu, Chao Ma, Chunyang Gao, Huihua Yang, Tao Xu, Rushi Lan and Xiaonan Luo, "Cost-Sensitive Collaborative Representation Based Classification via Probability Estimation Addressing the Class Imbalance Problem", 2018

[17] Chongsheng Zhang, Jingjun Bi, Shixin Xu, Enislay Ramentol, Gaojuan Fan , Baojun Qiao, Hamido Fujita, "Multi-Imbalance: an open-source software for multi-class imbalance learning", 2019

[18] Salvador García , Zhong-Liang Zhang, Abdulrahman Altalhi, Saleh Alshomrani, Francisco Herrera, "Dynamic ensemble selection for multi-class imbalanced datasets", 2018

[19] Dariusz Brzezinski , Jerzy Stefanowski, Robert Susmaga, Izabela Szczęch, "Visual-Based Analysis of Classification Measures and their Properties for Class Imbalanced Problems", 2019

[20] Sergio González, Salvador García , Sheng-Tun Li, Francisco Herrera, "Chain based sampling for monotonic imbalanced classification", 2018

[21] Sebastián Maldonado , Julio López, "Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification", 2018

[22] Ludmila I. Kuncheva, Álvar Arnaiz-González, José-Francisco Díez-Pastor, and Iain A. D. Gunn, "Instance Selection Improves Geometric Mean Accuracy: A Study on Imbalanced Data Classification", 2018

[23] Shiven Sharma, Anil Somayaji, Nathalie Japkowicz, "Learning over subconcepts: Strategies for 1-class classification", 2017

[24] Bartosz Krawczyk , Mikel Galar , Michał Wo źniak, Humberto Bustince , Francisco Herrera, "Dynamic ensemble selection for multi-class classification with one-class classifiers", 2018

[25] Mateusz Lango, Jerzy Stefanowski, "Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data", 2018

[26] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

[27] https://www.datascience.com/blog/imbalanced-data

[28] https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[29] https://elitedatascience.com/imbalanced-classes

[30] https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/