*Urban Studies: Border and Mobility – Kerr et al. (Eds)*
*© 2019 Taylor & Francis Group, London, ISBN 978-1-138-58034-3*

# A Survey on Data Level Techniques - A Customer Churn Prediction Case Study

Gillala Rekha[1], Shaveta Malik[2], Amit Kumar Tyagi[3], V Krishna Reddy[4]
[1,4]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, India.
[2]Terna Engineering College, Navi Mumbai, Maharashtra, India.
[3]School of Computing Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai, Tamilnadu, India.
gillala.rekha@klh.edu.in, shavetamalik687@gmail.com, amitkrtyagi025@gmail.com, vkrishnareddy@kluniversity.in

**Abstract:** Customer churn prediction and retention is a major issue for various service based organizations and skewed data representation presents significant challenges for such problems. The class imbalance exists when the number of samples of one class is much lesser than the ones of the other classes. In machine learning, the data-level approaches are known to handle the class imbalance problem. In this paper, we comprehensively study the performance of different machine learning techniques in churn prediction with class imbalance. The analysis of the research literature has focus on the key role of Data Imbalance Problems in classification, handling the imbalanced data and data-level techniques used to overcome the skewed distribution. Finally, we uncover number of research implications and upcoming directions for regular, big data and deep machine learning.

***Keywords:*** *Class Imbalance Problem, Class Imbalance Problem, Class-Distributions, Data-Level Technique, Sampling, Classification.*

## 1. Introduction

Churn prediction is emerging as a great influential field in the category of customers. Companies are investing a huge amount of capital in the process of holding the customers, i.e. , preventing them from a churn. Every company is keen on attracting new customers as it is directly reflects to the profits of the company. Apart retention of the existing customers is also important. So, Customer Relationship Management (CRM) is one of the broadest strategies to improve the relations with the customers. It is broadly acknowledged and is being used today in the fields like telecommunication, e-commerce etc [26][29]. Therefore, adopting machine learning models that are able to predict customer churn can effectively help in customer retention campaigns and maximizing the profit.

Generally, machine learning or data mining algorithms assume an equal class distribution for the data. However, this may not be true for predicting customer churn data. The distribution of the data is skewed in nature wherein one class (negative/majority/non-churn) is represented with a high number of instances than the other ones (positive/minority/churn). For learning algorithms, this leads to great difficulty, as they are biased towards the majority class. The concept of designing a smart system for handling skewed distribution to overcome the bias is recognized as learning from unfair data [7]. Problem with skewed data is commonly addressed by the research community in the last two decades. The imbalanced data classification has drawn significant attention from academia, and industry. Many methods were developed for handling imbalanced data for Customer Churn Prediction, focusing on balancing the imbalanced data using pre-processing techniques or modifying the existing classifiers. In general, traditional ML systems assume that the training datasets are fully-balanced with an equal misclassification error cost associated to each of the class [1].

Voluntary and involuntary churns are two main categories of churners [31]. Voluntary churners are those customers who make a decision to quit their services from the service providers. It is very difficult to decide/ determine these type of customers. The second type, involuntary churns are those customers whom the organizations decide to remove from the service. Therefore' this category includes people that are churned for fraud, non-payment and customers who don't use the phone. Voluntary churner is more difficult to determine; it occurs when a customer makes a decision to terminate his/her service with the provider.

The remaining part of this paper is structured as follows. The definition of customer churn us presented in section 2. The overview of class imbalance problems in classification is defined in Section 3. Section 4 presents the current research on Customer Churn Prediction at data-level. Section 5 addresses the issue on customer churn prediction using data level techniques. Further, research implication and future directions are presented in Section 6. Section 7 presents the challenges pertaining to class imbalance distribution. Further, section 8 provides several future research directions towards class imbalanced datasets/ problem. Finally, we make concluding remarks in Section 9.

## 2. Customer Churn

The term Customer churn is mostly used in telecommunication industry. It denotes the movement of customers from one service provider to another. The term "churn" means the movement of the customer to new service provider. The reasons for this movement may be unhappiness with service quality, unpleasant plans, high cost, etc. Furthermore the customer can quit the service due to financial problems, change of geographical location and sometimes the company may withdraw the service due to policy reasons. Customer churn prediction methods help the service providers in identifying the customer who likely to move to different service provider in near future. Poel et al., [27] focused on four sets of data variables like, customer behavior, customer perceptions, and customer demographics and macro environment in customer retention. By recognizing

the customer behavior regarding the service utilization, the number of calls, usage of network for data exchange, etc the service provider can decide the churn customers.

For an effective customer prediction, following steps need be used:

*File reader -> Read Data-> Data Manipulation -> Visual data Exploration -> Data Analysis-> Scoring->Drill down and Report*

Note that above structure (process) for predicting customer churn has been discussed with KNIME simulator. Client discernments are distinguished as the manner in which a client get or stop the service and can be determined with client reviews and incorporate information connect by and large happiness, nature of service, issue understanding, fulfillment with issue dealing with, intrigue given, area comfort, picture or notoriety of the organization, client view of reliance to the seller, and so forth. Client statistic incorporates age, sex, training, societal position, land information are likewise utilized for stir estimation

## 3. Imbalanced Data Classification Problem

In general, the Imbalanced dataset problem comes in the category of classification, when the sum of instance of one class is lesser than the instance of another class.. Furthermore the class of interest is one with smaller number of instances [4]. In real-time applications, this problem is of great interest. Most often, conventional classifiers have a bias towards the classes with greater number of instances. In turn, the “minority class” is usually ignored by treating them as noise. In this way, minority class samples are most often misclassified than other classes. The learning task does not hinder only by skewed data distribution but also series of problem like small size samples, overlapping between classes and small disjuncts may occur. In Figure 1, shows the examples of the three different kinds of imbalanced data distribution.

a) Small size sample/ class imbalanced data distribution: It refer to unequal or skewed distribution problem wherein not all classes for given dataset are represented similarly. The high Imbalanced Ratio (IR) may lead to poor learning, resultant in whole bias for the majority class [5].
b) Overlapping between classes: In the presence of overlap between the majority/ negative and minority/ positive classes, the classifier tends to imperfectly categorize the minority instances [9]. Hence, combination of overlapping between the classes with high IR usually outcomes in high misclassification rate for the minority class samples.
c) Small disjuncts: When the class consisted of smaller sub concept than the small disjucts in a dataset occurs and it increases the complexity for the training algorithm.

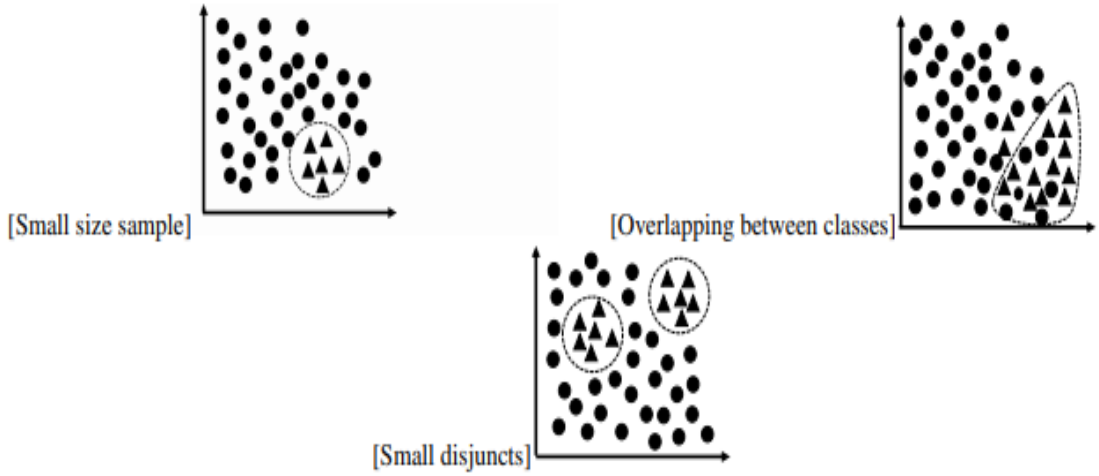


Fig. 1. Example of class distribution for two-class imbalance problems

### 3.1 Different Approaches in Class Imbalance Problems

The “class imbalance” problems in classification occur when there are significantly lesser samples in one class compared to other class. In current years, “class imbalance” problem has attracted the attention of researchers. Broadly, class imbalance problem can be addressed at three levels.

- Data level methods: To stabilize the class distribution the sample dataset is changed at data level. These methods apply pre-processing method to balance the “skewed distribution” in data.
- Algorithm level methods: These methods adapt the present learning algorithms to lessen the bias to negative classes and familiarise them to classify the data with “skewed distribution”. These techniques provide cost sensitive learning by taking misclassification costs into consideration.
- Ensemble methods: These methods use ensembles of classifiers and also known as ensemble methods along with data-level techniques. These methods increase the accuracy by training multiple classifiers and combine their output.

Hence in this work, we discuss different data level techniques applied to solve Customer Churn Prediction. Now, next section will discuss about the different techniques applied to solve customer churn prediction using data level techniques.

## 4. Data-Level Techniques for Class Imbalanced Problem - Customer Churn Prediction

In data-level approach, the sample dataset is modified to balance the class distribution. The foremost aim is to maintain equality in the class distribution for the datasets using sampling methods such as over-sampling, under-sampling and combination of both. The oversampling and under-sampling techniques are the two popular techniques in sampling-based classification to address the imbalanced datasets. In the oversampling technique, some samples are added to the minority class to make it balanced when very less information is available for minority class samples. In the under-sampling technique, some samples of the majority class are eliminated to make the dataset balanced. Apart from above, the hybrid techniques usually come with a combination of both over and under-sampling methods. Figure 2 categorizes different approaches applied at data-level to address the class imbalance problem.

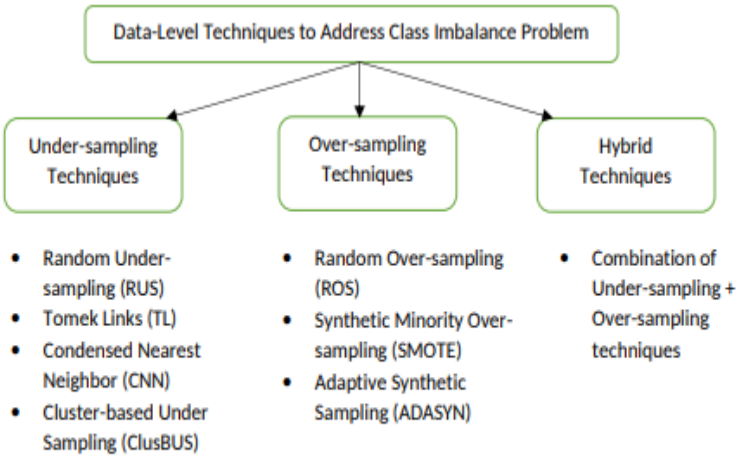


Fig. 2. Different data-level techniques to address the class imbalance problem.

The data-level techniques are also called as data pre-processing methods. These techniques are independent from classification stage. In the literature, many techniques have been proposed to handle skew distribution using sampling methods. Empirically it is proved that application of data-level technique to balance the skewed distribution before trained on a classifier yield better solution. There are different oversampling and under-sampling techniques that exist in literature such as Random Under-Sampling (RUS) [11], Random Over-Sampling (ROS) [6], Tomek links [10], Synthetic Minority Over-Sampling Technique (SMOTE) [3], One-Sided Selection (OSS), Condensed Nearest Neighbor Rule (CNN), Neighborhood CLeaning rule (NCL) [8], SMOTE+Tomek links [2] etc. Next, we will outline the different techniques proposed to handle skewed data distribution in Customer Churn Prediction in Table 1.

**Table 1: Different Techniques used, Algorithms Applied and Conclusion Drawn**

| Citation | Technique Used | Algorithm Used | Conclusion drawn |
|---|---|---|---|
| [12] | Mega-Trend Diffusion Function (MTDF), Synthetic Minority Oversampling TEchnique (SMOTE), Adaptive Synthetic Sampling Approach (ADASYN), K-Nearest Neighbor (K-NN), Majority Weighted Minority Over- Sampling Technique (MWSMOTE), and Immune Centroids Oversampling Technique. | Genetic Algorithm (GA). | Overall performance of MTDF and rules-generation based on genetic algorithms performed better when compared with other oversampling methods and rule-generation algorithms. |
| [13] | SMOTE and MTDF. | Rule generation algorithms like Exhaustive, Genetic, Covering, and LEM2. | The predictive performance of both oversampling techniques and rules generation algorithms was outstanding. |
| [14] | Over-Sampling (OS), Under-Sampling (US) and SMOTE. | Random Forest (RF) | Combination of data level approaches with classifiers will maximize the turnover with the minimum overheads for customer retention and help reduce customer churn rates. |

| | | | |
|---|---|---|---|
| [15] | Particle Swarm Optimization (PSO) based under-sampling method. | Genetic Programming (GP)-AdaBoost (GPAB) | The proposed GPAB yields better AUC over other evaluation metrics. |
| [16] | SMOTE. | Classification and Regression Trees(CART), Bagged CART and Partial Decision Trees(PART). | SMOTE along with learning algorithms provide better results for predicting customer churn. |
| [17] | combination of simple under sampling and SMOTE. | Weighted Random Forest (WRF). | Combined sampling and WRF produced better a prediction model than stand alone classifiers. |
| [18] | US, OS, US+OS and SMOTE. | fuzzyARTMAP. | fuzzyARTMAP performed exceedingly well using sensitivity metric. |
| [19] | One Class support vector machine (OCSVM) based undersampling method. | Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Probabilistic Neural Network (PNN) and Group Method of Data Handling (GMDH). | The author specified that DT over other classifiers along with “if-then" rules, achieved high AUC. |
| [20] | Local Principal Component Analysis (PCA). | C4.5 -DT, SVM, LR and the Naive Bayes (NB) | Local PCA along with Smote outperformed Linear regression and Standard PCA data generation techniques. |
| [21] | ADASYN | Backpropagation algorithm (BP) | The result of the study with proposed scheme shown better accuracy and F1-score. |

In [22], the author applied correlation analysis among the different classifier’s accuracy and certainty of its prediction. In [23], the author preformed comprehensive study on the performance of decision tree in churn prediction with class imbalance. They conclude that the findings provided will act as useful guideline for usage of decision tree in churn prediction. In [24], the author implemented exploratory data analysis and feature engineering on public telecom dataset using different classification techniques like Naïve Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Trees. The result expressed that the best classifier is Gradient Boosted Trees. Ahmed et al [25] presented a study on deeper understanding of customer churn for telecom industry using various machine learning algorithms.

Hence, this section discusses about different techniques used, algorithms applied and conclusion drawn in past for solving customer churn prediction .Now, next section discuss about various machine learning algorithms used.

## 5. Machine Learning Algorithms used for Customer Churn Prediction

In this section, we briefly present the popular machine learning techniques used for customer churn prediction. The most popular algorithms used by the research community in the past decade are Decision Trees Learning, Support Vector Machine, Artificial Neural Network, Naïve Bayes and Regression analysis. These algorithms are considered due to their efficiency, reliability and popularity. [29,30]

5.1 Decision Tree Learning: Decision Trees (DT) is a classification model. The tree-shaped structure of DT represents set of decisions for generating classification rules for a specific data. There are different variations such as C4.5, Classification and Regression Trees (CART). In these tree structures, the class labels are represented as leaf nodes and branches represents the outcomes of features. DT presents good performance accuracy when applied to customer churn problem.

5.2 Support Vector Machines: Support Vector Machines (SVM), also known as Support Vector Networks, introduced by Boser, Guyon, and Vapnik [33], are supervised learning models. SVM uses support vectors and analyze the given data to derive useful patterns. It is used for both classification and regression analysis. SVM employ kernel functions for improving the performance. Selecting the best kernels or combination of different kernels is still an open research. In the churn prediction problem, SVM outperform DT, depending mainly on the type of data and data transformation that takes place among them.

5.3 Artificial Neural Network: Artificial Neural Networks (ANN) is a popular approach to address typical classification problems, such as the churn prediction problem. Neural networks works on neurons associated with weights for each neuron. Different topologies have been defined to make the learning system work for varied problems. One of the most popular supervised model built using ANN is Back-Propagation algorithm (BPN). BPN is a feed-forward model with supervised learning.

5.4 Naive Bayes: A Bayes classifier is a probabilistic classification algorithm. It is based on Bayes’ theorem. It is an independent feature model, with prior and posteriori probability estimates. A Naive Bayes (NB) classifier assumes that the incidence (or nonexistence) of a particular feature of a class (i.e., customer churn) is unrelated

to the incidence (or nonexistence) of any other feature. The NB classifier achieved good results on the churn prediction problem for the wireless telecommunications industry.

5.5 Regression Analysis: Regression analysis is based on statistical model. It is used for estimating the relationships among features. It includes many techniques for developing several variables. Regression analysis is mainly used to find the relationship between a dependent variable and one or more independent variables. In terms of customer churning, logistic Regression mainly been used. It is a type of probabilistic statistical model used to find a binary prediction of a categorical variable (e.g. customer churn) which depends on one or more predictor variables (e.g. customers' features).

Hence, this section discusses several exiting algorithms which are useful in prediction churn prediction. Now, next section will discuss several practices related to imbalance data-sets in many applications (in this smart era).

## 6. Current Practices and Applications in Smart Era

Organizations today have huge amount of data (structured and unstructured form) that provides further insights about the customer retention. This data can be helpful in identifying who, why and when the customer churn, as well as target customers and market tricks. Till data, research went on in prediction of churn risk based on the customer activity and demography. The researchers employed mostly traditional approaches such as probability model, regression models and vulnerability models. Recent advancement in technology lead to data collection from various sources such social media, online chats, and web. In addition, applications of machine learning algorithms have opened a better opportunity for customer churn prediction and retention research [28]. Advances in machine learning domain enable researchers to extract useful information from unstructured data (audio and video, images). We believe that data on social connections will gather more attention. The main concern is to understand why the customer is at risk and whom to target. Apart textual data will also provide insight for customer churning. The development various predictive models is useful in building the predictive models to find the risky customer, to analyze the market trends, to identify the target customers and also the strategy campaign targets. The more advanced approaches like deep learning based on neural networks can learn the customer probability and retention modeling. Apart, feature and variable selection techniques like dimensionality reduction form a major development in the field of machine learning for customer churn prediction.

Hence, this section discusses about an introduction to customer churn pertaining to an organization. Now, next section will include several challenges in imbalanced data-sets/ solving class imbalance problem.

## 7. Challenges

Napierala and Stefanowski [32] proposed different method to analyze the minority samples by assigning it to predefined categories such as safe, borderline, rare and outliers. Such methods help in understanding the difficulties present in the data. Hence, some challenges are included here as:

a) As a future direction, it is important to propose new classification algorithms that incorporate the different difficulties in the data. Apart, while designing the classifier, attention should be needed towards individual minority samples. Another important issue is extreme class imbalance problems. The extreme imbalanced data sets exist in most of the real-world problems such as fraud detection with Imbalance Ratio (IR) approximately 1:3000. This poses a great challenge for classification algorithms to train on such extreme datasets. Third challenge is inefficient features extraction for some problems such as protein data, online transaction data. It is very much important for the classifier to be trained on such high- dimensional and sparse feature set.

b) Another way to solve class imbalance problem is by modifying the learning algorithm. However, a major drawback of such learning models gives much importance to minority samples, thus increasing the majority class misclassification. A technique needs to be proposed to select only uncertain samples and adjust the output accordingly.

c) Recently, ensemble learning became the most popular techniques for handling class imbalance. Algorithms like Bagging, Boosting, stacking, and Random Forests were robust in handling data difficulties. Ensemble learning along with sampling techniques provides better performance to handle the skewed distribution of data. The main drawback is diversity among majority and minority class. There is no proper indication of how large the ensemble should be constructed as their size is selected arbitrarily.

d) Another problem is handling of multi-class imbalanced classification. The multi-class imbalance occurs when more than two classes with one majority class and multiple minority class exist. A deeper insight in handling multi-class imbalanced problems is needed.

e) Data pre-processing technique is highly importance in balancing the imbalance datasets as there are independent of classifiers. The possible difficulties appear in the data such as class overlapping, noise and small disjunct. Therefore, efficient data cleaning and sampling techniques are needed to balance the data. For multi class imbalance problems, efficient sampling techniques need to be proposed.

f) In multi class imbalance learning need special care while applying sampling techniques. Researchers should focus on developing algorithms which are robust in handling such skew distributions.

g) In ensemble learning algorithms such as bagging and boosting ,there may be different level of uncertainty while sampling the data into bags. There may be a high probability of consisting samples from the same class within a single bag. The need for proper probability distribution techniques such as normal, binomial distribution can be used to check the balance distribution in each bag. However, many difficulties may arise due to data distribution in each of the bags and also each bag may contain certain amount of noise which makes the classifier to perform poorly. So,

efficient techniques need to be proposed in handling the size of the bags and the distribution of samples into each bags.

h) Another important and yet popular challenge in class imbalance is learning from continuous data. The process of learning from the continuous data is called data streaming. The need for active learning algorithms to address data streaming issue is still at infant. The general open issue will be based on sampling the streaming data and classifying it.

i) In last, extraction of efficient features and instance is also of major concern. Real time data such as bank data or genomic data are essentially have high-dimensional and sparse feature. The development of new approaches for high dimensional data is much needed, that will allow at the same time for an efficient processing and boosting discrimination of the minority class. Another interesting direction is to investigate the possibilities of using decomposition-based solutions. Note that a lot of to overcome above issues and challenges, several techniques have been proposed (used at during the pre-processing stage), in that, most popular one is resampling, which includes under-sampling and oversampling techniques.

Hence, this section includes/ discusses several serious challenges available in current towards class imbalance problem/ data-sets. Now next section will discuss several future research directions in customer churn prediction.

## 8. Future Research Directions

The growing interest and capable performance have made data mining and machine learning techniques to be potential in customer churn prediction. Although, comprehensive in machine learning and data mining application in customer churn prediction, this paper presents some insight into the challenges and future direction. Firstly, the customer churn data comes from various source and in large volume, efficient machine learning algorithms are required to handle unstructured and click stream data. Next, the nature of the data is be highly imbalance proper care to be taken to avoid the problem of losing information in the case of under-sampling and the overfitting problem at oversampling. Meanwhile, new technologies in big data can also be used and development of cloud computing and internet of things also influence the embedded analytics and the development of dynamic big data analytics networks. This paper aims at identifying the different sampling techniques and algorithms used for customer churn prediction. Once churns are identified the organization can take action for preventing the churns. The company can identify such churns and put efforts to retain the customers.

Table 2: Churn Prediction Categories

| | Actual Churners | Actual Non Churners |
|---|---|---|
| Predicted Churners | TP (True Positive) | FP (False Positive) |
| Predicted Non Churners | FN (False Negative) | TN (True Negative) |

Table 2 shows the churn prediction categories in the form of confusion matrix. The confusion matrix with true positive, true negative, false positive and false negative helps in evaluation process for the churn prediction data. The common evaluation metrics are accuracy, precision, recall, sensitivity, specificity, and Area Under Curve (AUC). As discussed in [34], future researchers/ research community should consider the following research directions (for providing solutions to imbalanced problems):

a. A better understanding on the nature and structure of samples in minority classes may provide insight to the learning difficulties in handle it more efficiently.
b. New methods or algorithms need to be developed for handling multi-class imbalanced datasets by taking varies relationships between classes into consideration.
c. New solutions need to be proposed for multi-instance and multi-label learning for handling skewed distribution of data.
d. Efficient clustering methods need to be introduced for class overlapping problems to partition the datasets and to select proper evaluation model in such scenarios.
e. The deeper analysis of individual properties for each rare class may provide an intuition in handling the class imbalance problem in very efficient way.
f. The effective way of handling not only structured data but also studying of unstructured and streaming data is needed and researcher must focus in this area.
g. As there is huge volume of data generated using IoT and sensor devices, potential tools are needed and deep insight in analyzing the big data is a hot topic at present.

Hence, we need to overcome and look forward to above discussed future research directions. This section discusses several future research directions with related to imbalanced data-set. Now, next section will conclude this work in brief.

## 9. Conclusion

In this paper we discussed about imbalance data problem and number of methods have been used to develop to solve the problem of imbalance in Customer Churn Prediction and various techniques to enhance the performance with respect to the class imbalance .Moreover the method is divided into three parts which is depends on their data level technique and the

discuss about the class imbalance problem .The major challenges is the big data and the large amount of data. Previously data mining methods were used but those are not up to the mark with the new requirements forced by big data. Since the diversity and accuracy of the composed evidence big data is also pretentious by this kind of twisted circulation. So, we present the different techniques proposed to handle class imbalance big data at pre-processing stage. Finally, we also present the challenge that need to be addressed for big imbalance data. The paper review will help the scientist, researchers and useful in the new research track with respect to the expansion and valuation.

**Acknowledgements**

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

**References**

[1] Anand, R., Mehrotra, K.G., Mohan, C.K., Ranka, S., 1993. An improved algorithm for neural network classification of imbalanced training sets. IEEE Transactions on Neural Networks 4, 962–969.
[2] Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter 6, 20–29.
[3] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.
[4] Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter 6, 1–6.
[5] Das, S., Datta, S., Chaudhuri, B.B., 2018. Handling data irregularities in classification: Foundations, trends, and future challenges. Pattern Recognition .
[6] Fan, X., Tang, K., 2010. Enhanced maximum auc linear classifier, in: Proceedings of the 7th international conference on Fuzzy systems and knowledge discovery (FSKD), IEEE. pp. 1540–1544.
[7] Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. Intelligent data analysis 6, 429–449.
[8] Laurikkala, J., 2001. Improving identification of difficult small classes by balancing class distribution. Artificial Intelligence in Medicine, 63–66.
[9] Prati, R.C., Batista, G.E., Monard, M.C., 2004. Class imbalances versus class overlapping: an analysis of a learning system behavior, in: Mexican international conference on artificial intelligence, Springer. pp. 312–321.
[10] Tomek, I., 1976. Two modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics 6, 769–772.
[11] Xiaoying, X., Sheng, F., 2012. A synthesized sampling approach for improving the prediction of imbalanced classification, in: Proceedings of the IEEE 3rd International Conference on Software Engineering and Service Science (ICSESS), IEEE. pp. 615–619.
[12] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A. and Hussain, A., 2016. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, *4*, pp.7940-7957.
[13] Amin, A., Rahim, F., Ali, I., Khan, C. and Anwar, S., 2015. A comparison of two oversampling techniques (smote vs mtdf) for handling class imbalance problem: A case study of customer churn prediction. In *New Contributions in Information Systems and Technologies* (pp. 215-225). Springer, Cham.
[14] Gui, C., 2017. Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. *Artif. Intell. Research*, *6*(2), p.93.
[15] Idris, A., Iftikhar, A. and ur Rehman, Z., 2017. Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. *Cluster Computing*, pp.1-15.
[16] Mishra, K. and Rani, R., 2017, August. Churn prediction in telecommunication using machine learning. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 2252-2257). IEEE.
[17] Effendy, V. and Baizal, Z.A., 2014, May. Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. In *2014 2nd International Conference on Information and Communication Technology (ICoICT)* (pp. 325-330). IEEE.
[18] Naveen, N., Ravi, V. and Kumar, D.A., 2009. Application of fuzzyARTMAP for churn prediction in bank credit cards. *International Journal of Information and Decision Sciences*, *1*(4), pp.428-444.
[19] Sundarkumar, G.G., Ravi, V. and Siddeshwar, V., 2015, December. One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-7). IEEE.
[20] Sato, T., Huang, B.Q., Huang, Y. and Kechadi, M.T., 2010, August. Local PCA regression for missing data estimation in telecommunication dataset. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 668-673). Springer, Berlin, Heidelberg.
[21] Aditsania, A. and Saonard, A.L., 2017, October. Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm. In *2017 3rd International Conference on Science in Information Technology (ICSITech)* (pp. 533-536). IEEE.
[22] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J. and Anwar, S., 2019. Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*, pp.290-301.
[23] Zhu, B., Xie, G., Yuan, Y. and Duan, Y., 2018, May. Investigating Decision Tree in Churn Prediction with Class Imbalance. In *Proceedings of the International Conference on Data Processing and Applications* (pp. 11-15). ACM.

[24] Halibas, A.S., Matthew, A.C., Pillai, I.G., Reazol, J.H., Delvo, E.G. and Reazol, L.B., 2019, January. Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling. In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1-7). IEEE.
[25] Ahmed, A. and Linen, D.M., 2017, January. A review and analysis of churn prediction methods for customer retention in telecom industries. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1-7). IEEE.
[26] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M. and Abbasi, U., 2014. Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, *24*, pp.994-1012.
[27] Van den Poel, D. and Lariviere, B., 2004. Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, *157*(1), pp.196-217.
[28] Amit Kumar Tyagi, Rekha, "Challenges of Applying Deep Learning in Real-World Applications", Book-Challenges and Applications for Implementing Machine Learning in Computer Vision, IGI Global, 2020.
[29] Gillala Rekha et al. "A Wide Scale Classification of Class Imbalance Problem and its Solutions - A Systematic Literature Review" Journal of Computer Science 2019, 15 (7): 886.929.
[30] G Rekha, Amit Kumar Tyagi, "Necessary Information to Know to Solve Class Imbalance Problem: From a User's. Perspective", ICRIC 2019, March 2019.
[31] Kraljević, G. and Gotovac, S., 2010. Modeling data mining applications for prediction of prepaid churn in telecommunication services. *Automatika*, *51*(3), pp.275-283.
[32] Napierala, K. and Stefanowski, J., 2016. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, *46*(3), pp.563-597.
[33] Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.
[34] Raghava Lavanya, G.Rekha, Amit Kumar Tyagi, Niharika Sakruti, "Class Imbalanced Data: Open Issues and Future Research Directions", ICSCT 2019.