ISSN (Online) 2395-2717

International Journal of Engineering Research in Electrical and Electronic Engineering (IJEREEE)

Vol 8, Issue 6, June 2021

Data Sampling Techniques for Network Analysis through Big Data

[1] S.S.Raja Kumari, [2] Dr. CH.Nagaraju

[1] Associate Professor, Department of CSE, St.Johns College of Engineering and Technology, Yerrakota, Yemminganur, Kurnool District, Andhra Pradesh, India
ORCID ID- 0009-0007-7918-0970
[2] Professor and HOD, Department of ECE, Annanamacharya Institute of Technology and Sciences (AITS), Rajampet, Andhra Pradesh, India
ORCID ID- 0000-0001-9178-6448
Corresponding Author Email: [1] ssrajakumari2009@gmail.com, [2] chrajuaits@gmail.com

Abstract— ***Data sampling in big data networking analysis is essential for differentiating associations and patterns within large data sets. There are plenty of various methods to gather an investigation's sample, all of which have a unique set of advantages and disadvantages to avoid errors and biases. Adaptive and resampling techniques assist in overcoming the challenges of bias, error, and complication in the sampling procedure. Practical uses for inverse sampling to decrease class differences in machine learning are addressed.***

Keywords: Data Sampling, Big Data Analysis, Sampling Techniques, Class Imbalance, Inverse Sampling, Network Analysis, Machine Learning.

I. INTRODUCTION

A statistical analysis technique that is data sampling is used to select, adapt, and examine a representative "sample points of data" in order "to identify trends and patterns" in the larger data set under study. Having a restricted, adequate quantity of data about a statistical population makes it possible for "data scientists", "predictive modelers", and various "other data analysts" to create mathematical models faster while still generating reliable results [1]. A common method of statistics for many kinds of uses, such as viewpoints, web analytics, and political surveys, is data sampling. Moreover, without having to collect and analyze data from each member of the general population, forecasts about the wider population can be generated with a certain amount of confidence via a data sample. Finding efficient methods for obtaining representative sets is the objective of studies on data sampling methods for large data-driven networks.

The main objectives of the study

- To assess the current landscape of network analysis in the context of big data
- To develop a comprehensive understanding of the trade-offs between different data sampling methods
- To evaluate the performance and reliability of the developed framework through empirical experiments

II. LITERATURE REVIEW

Types of sampling techniques in data analytics

In big data analysis, when analyzing and conserving an entire data set can be costly and time-consuming, sampling can be especially useful. However, there are multiple disadvantages and limitations to sampling that analysts have to consider. Various methods of sampling can be employed by analysts based on the goal, the data's access, and its characteristics. For instance, stratified sampling can be useful for different information with different subgroups, whereas an obvious random sample performs best for homogenous and uniformly distributed data. For big and scattered data, cluster sampling performs well, whereas systematic sampling is simple to employ for consecutive and ordered data [2]. Each approach could have disadvantages of its own, like introducing elements of bias or error in the occurrence that the data set is not representative of the general population.

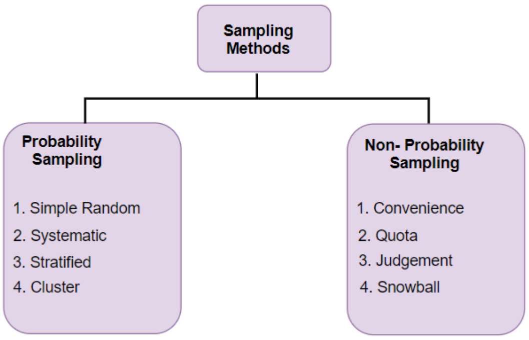

Figure 1: Types of sampling techniques in data analytics (Source: [2])

Sampling challenges that are faced in big data's data analysis

Sampling bias, sample error, and sample size are a few of the issues that analysts might face when employing samples for big data analysis. Once a sample is not representative of the population, sampling bias develops and may result in inaccurate or deceptive findings. Analysts have to make sure the sample size, method, and source are suitable for the study in question and the nature of the data, therefore, in order to avert this [3]. The difference between the parameter for population and sample statistic as a consequence of random or intentional sampling variations is referred to as sampling error. Analysts must determine the tolerance of error and credibility interval, apply suitable sampling strategies, and conduct sensitivity and robustness evaluations in order to minimize this error. The challenge or expense of obtaining and evaluating an example from a sizable, complex data set that may have numerous formats, sources, frameworks, and characteristics is referred to as sampling complexity [4]. Analysts have to use platforms, tools, and systems that are able to process a great deal of data and enable sample processes in order to address this issue.

Solution to mitigate challenges of data analysis

Utilizing adaptive sampling, resampling, and population integration, researchers may improve the accuracy and precision of their big data studies. In adaptive sampling, data from previous samples is utilized to improve both the technique of sampling and the sample size. In order to figure out the pattern of sampling or perform statistical tests, resampling generates multiple samples from the initial sample or community [5]. As many samples or methods of sampling are combined, a more comprehensive and precise sample can be produced. These techniques have a chance to enhance sampling reliability and validity, increase coverage and diversity, streamline the process, and reduce complexity.

III. METHODS

Positivism research philosophy has been chosen for conducting the study. Instead of being rigid about a specific theoretical framework, researchers who are pragmatic are occupied with getting the job done. A positivism approach allows for versatility in fulfilling the goals of the study which is particularly crucial in a difficult field including big data network analysis [6]. It has been chosen to use a sequential *exploratory design* for the study. In order to examine the circumstances and problems of using huge amounts of information in network studies, this approach begins with the collection and analysis of qualitative data. Furthermore, concrete solutions to these issues could be created with the help of this qualitative study. The research will employ a *secondary qualitative* data analysis method. The data has been gathered from secondary resources, including scholarly publications, news stories, and online databases. The collected data has been presented through thematic data analysis.

IV. RESULT

Integrating survey data into big data analysis utilizing inverse sampling

Assuming there is a systematic difference between the pattern of distribution of the research variables in the vast data sample and the distribution in the population of interest, then the large data sample is not representative of the population of interest [7]. The key concern is how well researchers are able to compensate for the choice bias by using additional variables that have been excluded from the huge sample. In this section, researchers consider an unusual inverse sampling technique for dealing with this issue.

$$\hat{\theta}_{B1} = \frac{\sum_{i\in B} \frac{f(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i|\delta_i=1)} \frac{f(y_i|\boldsymbol{x}_i)}{f(y_i|\boldsymbol{x}_i,\delta_i=1)} y_i}{\sum_{i\in B} \frac{f(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i|\delta_i=1)} \frac{f(y_i|\boldsymbol{x}_i)}{f(y_i|\boldsymbol{x}_i,\delta_i=1)}}$$

Inverse sampling can be considered an instance of two-stage sampling. The large data sample utilized during the first phase has bias in selection as it was gathered in the initial phase. In order to compensate for the possibility of selection bias created by the enormous information sample, researchers conducted another round of sampling. It is initially claimed that inverse sampling may be employed to turn an intricate sample into a single random sample. Researchers use the inverse sampling algorithm for particular classic designs, "such as stratified sampling". Researchers applied the inverse sampling idea to an allocated sample. In this section, they discuss how inverse sampling can be used to reduce selection bias in large data sets.

As significant weights are determined for every element in the large data sample, this step can be known as the weighting stage. The second stage entails choosing an area of the enormous amount of information to be analyzed with a probability of selection that fits its significance weights [8]. The indicated inverse sampling approach for eliminating selection bias needs data on the intended population from a different source, such as a census or a sample of probability, for the purpose of calculating an auxiliary variable x. To formally introduce the concept, researchers assume that the big data sample includes the variables xi and yi while the density of the median distribution of x can be calculated by the equation f.x/. Researchers have assumed that the second value of the auxiliary variable x is finite. They are interested in calculating "D.E.Y. from the big data sample B". For the vast data set B, they do not know the "first-order inclusion probability".

$$\hat{\theta}_{B1} = \frac{\sum_{i\in B} \frac{f(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i|\delta_i=1)} y_i}{\sum_{i\in B} \frac{f(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i|\delta_i=1)}} := \sum_{i\in B} w_{i1} y_i .$$

Big data and survey data can be linked with the assistance of survey data collection strategies. In this section, researchers assume two sources for information, A (survey data) and B (big data), each of which is susceptible to selection bias. The equation “n=NB D o.1/”, where n is the number of samples of A, serves under the presumption that data on item x is available from survey data and data on item y can be obtained from big data. They are interested in calculating the general mean YNN by combining two data sets.

Impact of Class Imbalance on Machine Learning and Sampling Techniques

The issue of class imbalance has an important effect on machine learning and techniques for deep learning. The last term refers to problems that arise when a single class in the data set has fewer samples than another group, resulting in a substantial decrease in the accuracy of the classifier. Many observations have been published in the scientific community about studies on this problem, with particular emphasis on data sampling methods including “random over-sampling (ROS)”, which replicates data from the minority class, and “random under-sampling (RUS)”, which eliminates samples from the entire class. These methods introduce an offset into the classification algorithm, which adjusts for the asymmetry between the categories.

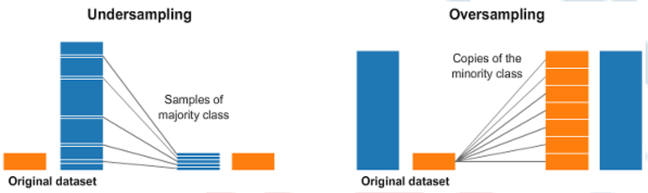

Figure 2: Class imbalance in ML
(Source: [9])

Replicating minority samples raises the likelihood of issues related to data sampling approaches, such as extended periods of training and overfitting. Once numerous samples have been removed from the majority of classes, the classifier could lose potentially pertinent data. Furthermore, a heuristic mechanism was developed along with additional “intelligent” methods for sampling. The heuristic over-sampling method that is “Synthetic Minority Over-sampling Technique (SMOTE)” has been created by the researcher that generates false samples of the disadvantaged subgroup by integrating adjacent real-world examples [9]. As a consequence of its achievement, many over-sampling algorithms have been created, and it has since become commonplace for the purpose of collecting data.

V. CONCLUSION

The study underlines the essential function of data sampling in big-data network analysis, tackling challenges such as class imbalances and biases. Inverse sampling and algorithmic heuristics like SMOTE are instances of new methods that help solve these issues. The research that makes use of actual approaches stresses the importance of resolving class imbalances to utilize machine learning. One potential method to increase accuracy and reduce bias in large-scale analysis of data is the inclusion of data from surveys using inverse sampling.

REFERENCES

[1] Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6(1), 1-16. https://doi.org/10.1186/s40537-019-0206-3

[2] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of big data, 7(1), 1-45. https://doi.org/10.1186/s40537-020-00369-8

[3] Aversa, J., Hernandez, T., & Doherty, S. (2021). Incorporating big data within retail organizations: A case study approach. Journal of retailing and consumer services, 60, 102447. https://www.sciencedirect.com/science/article/am/pii/S0969698921000138

[4] Chang, V. (2021). An ethical framework for big data and smart cities. Technological Forecasting and Social Change, 165, 120559. https://publications.aston.ac.uk/id/eprint/43893/1/VC_smartcity_transport_ethics_ver05F_accepted.pdf

[5] Wang, J., Yang, Y., Wang, T., Sherratt, R. S., & Zhang, J. (2020). Big data service architecture: a survey. Journal of Internet Technology, 21(2), 393-405. DOI: 10.3966/160792642020032102008

[6] Roy, R., & Uekusa, S. (2020). Collaborative autoethnography: “Self-reflection” as a timely alternative research approach during the global pandemic. Qualitative Research Journal, 20(4), 383-392. DOI 10.1108/QRJ-06-2020-0054

[7] Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An introductory review of deep learning for prediction models with big data. Frontiers in Artificial Intelligence, 3, 4. https://doi.org/10.3389/frai.2020.00004

[8] Saadoon, M., Hamid, S. H. A., Sofian, H., Altarturi, H. H., Azizul, Z. H., & Nasuha, N. (2022). Fault tolerance in big data storage and processing systems: A review on challenges and solutions. Ain Shams Engineering Journal, 13(2), 101538. https://doi.org/10.1016/j.asej.2021.06.024

[9] Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1560-1571.10.11591/ijeecs.v14.i3.pp1560-1571.