A COHERENT WAY OF SAMPLING - A RULE TO BUILD INSIGHTS

Sai Manasa Leburi
AI, Data Science & Machine Learning
Accenture
Hyderabad, India
sai.manasa.leburi@accenture.com

Anisha Sasi
AI, Data Science & Machine Learning
Accenture
Chennai, India
anisha.sasi@accenture.com

ABSTRACT- *No companies can serve without data these days. With huge quantities of data being generated every alternate from business deals, client logs and stakeholders, data acts as an imperative pillar that drives various firms. In this scenario, it is impossible to study the entire data population to draw conclusions that can help the business. Inferences drawn from the samples are intended to be generalized to the population, and sometimes to the future as well.*

Sampling is a statistical process that is concerned with the selection of the individual observation. It helps us to make statistical hypotheticals about the population.

Sampling ensures convenience, collection of intensive and exhaustive data, suitability in limited resources and better rapport and it permits you to draw decisions about very complex situations.

1. PROBLEM STATEMENT

In this digital era, data gets piled up in tera bytes that causes difficulty in doing analysis and may lead to pointless perceptivity, if not handled properly. During data analysis, analysts encounter various

challenges, including lack of storehouse space, computation time, lack of funding, backing in data collection etc.

Testing on all the data may be infeasible. You cannot go and measure every human being's weight and height if you wanted an estimate of the height and weight of people in your region. In computing, computational complexity associated with your algorithm may lead to bottlenecks as well.

As a result, experimenters can answer the utmost questions by testing a portion of data, rather than picking all the data population. Sampling permits you do your exploration briskly and at a lower cost.

2. PROPOSED SOLUTION

Before applying various analytics and data science related algorithms, data needs to be preprocessed. During the data preprocessing, data is modeled according to the requirement of individual data science algorithm, hence data preprocessing steps also vary.

When the data is tremendously large, the algorithms cannot perform well on the given data set and the computer memory is also employed further due to the larger data size. Similarly, it consumes time, the cost for data handling and related hardware requirements. In such scenarios, its necessary to pick a part of data that can represent the whole data population without losing any features and it is referred as sampling. This sampling can handle any data science model and is the introductory step of preprocessing.

The proposed solution focuses on efficacious method of data sampling. The method of sampling varies based on the type of analysis being performed.

IDENTIFY CLUSTER FEATURE SAMPLING

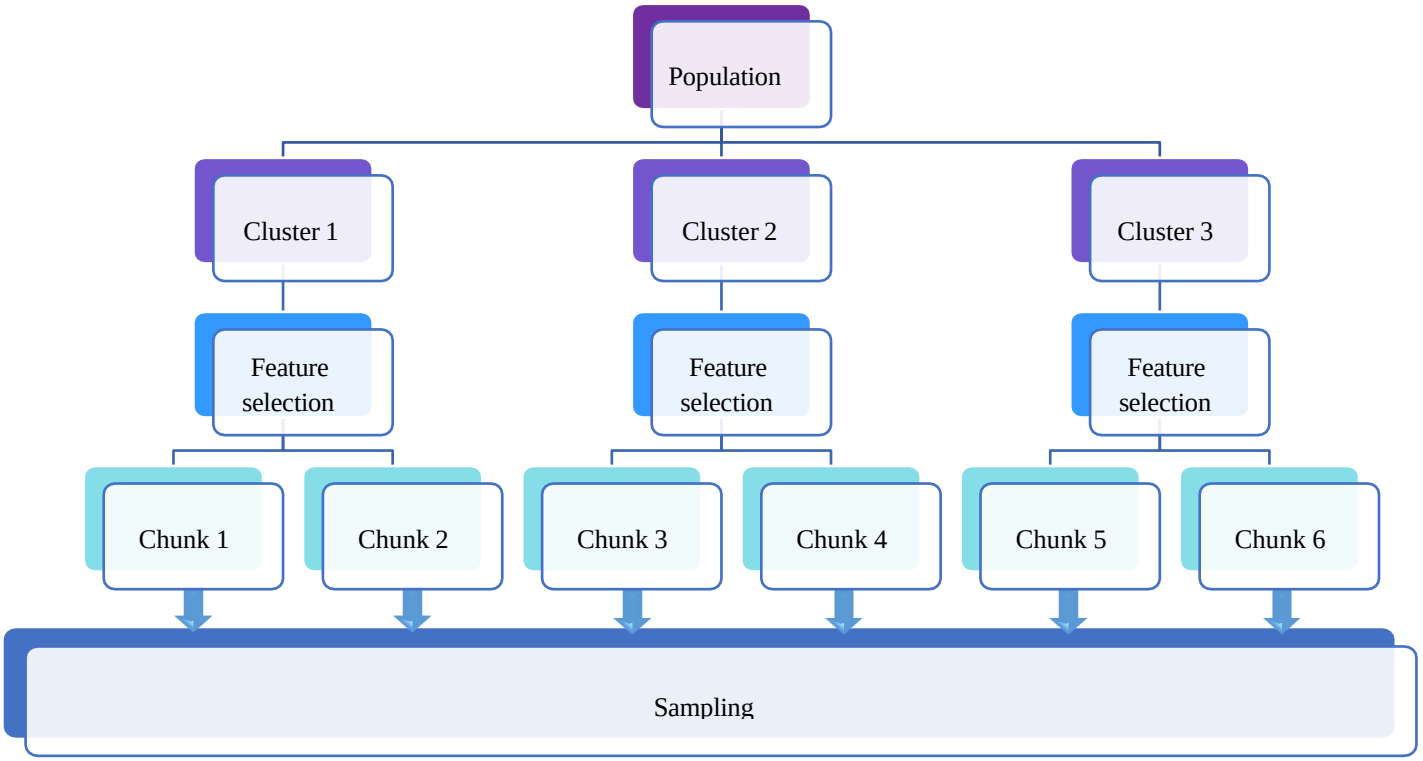

Figure 1: ICFS Flow Chart

The proposed sampling technique called Identify Cluster Feature Sampling (ICFS) helps in selecting a sample data that represent the whole population and, is one of the probability sampling techniques where entire population is divided into homogenous groups called smaller chunks and this chunk is used to complete the sampling process.

From the data population, we could identify the clusters based on different algorithms. After the identification of clusters, we could select the attributes or features that is essential for the model to perform its task. This attribute selection can be done with the help of feature engineering.

Now the clusters will have effective fields and through which we can divide the cluster into small chunks of data, which are similar in characteristics. Sampling can be performed on these chunks of data individually to get more accurate results.

2.1 Steps involved in ICFS:

1. Determine the population for your sampling.

   Population is created depending on the common observable features.

   Ex: A school can be considered as a population, where a variety of pupils data is composed of.

2. Identify the clusters that could evolve from the population.

   Clusters can be constructed from the entire population based on a generic shared behavior.

   Ex: If a school is taken as a population, then each classroom can be a cluster.

3. Attribute selection

   Once a cluster is formed, essential features can be extracted from the data. Through which an effective sampling can be performed.

   Ex: If a class is considered as a cluster, the best features could be gender, height, weight etc.

4. Create small chunks of data.

   With the help of identified attributes, smaller chunks/groups of data can be created.

   Ex: In a classroom, based on gender, height or weight students are divided further smaller groups

5. Specify the size of your sample based on the available infrastructure.

   It is imperative to define the ratio of the sample data, and this can proportionally present your whole data.

6. Randomly select the data from each chunk based on sampling proportion.

   Here we apply random sampling technique to select each data point from the respective chunk.

7. Merge all the chunk samples to one sample.

   Now we can merge all the chunks exemplifications into one and this can be used for a total population analysis.

2.2 Advantages of ICFS:

1. This is a systematic method of getting a population sample which has stronger exploration results.
2. This system is fair for all the data points, as the sample is selected from each group/chunk, and we could reduce certain level of biasness.
3. Data overlap between other groups is nullified, hence variations can be removed.
4. This is an affordable way to perform data analysis.

2.3 Experimental Data Details:

Below is the population and sample data that are used for this experiment.

TABLE NAME	RECORD COUNT	SIZE IN GB
Schema Population	340.37 MN	106.25
Schema Sampling	116 MN	14.41

Table 1: Experimental Data

	BEFORE SAMPLING	AFTER SAMPLING	COMMENTS
Data Size	**106.25 GB**	**14.41GB**	**Reduced by 85%**
Model Execution Time	**15 Days**	**1.5 Days**	**Reduced by 90%**
Cost Of Data Storage (Data stored in AWS S3. Refer section – Assumption)	**$2.5** (=106.25 * $ 0.023 Per GB)	**$0.33** (=14.41 * $ 0.023 Per Gb)	**Reduced by 87%**
Cost Of Server Utilization (AWS - 64 CPU, 512 Gb RAM)	**$1161.2** (=15 Days * $ 3.2256 Per Hr.)	**$116.2** (=1.5 Days * $ 3.2256 Per Hr.)	**Reduced by 89%**

Table 2: Result Comparison

From the Table 2, it is evident that the proposed sampling method is efficacious and can minimize the data size, execution time and cost of data storage and server utilization when huge data sets are handled.

3. FUTURE DIRECTIONS

Biasness can be induced into data while labelling, most of the time unintentionally, by humans. This can be because unconscious bias is present in humans. As this data teaches and trains the AI algorithm on how to analyse and give predictions, the output will have anomalies.

4. CONCLUSION

We have identified a better sampling process which can produce effective sampling regardless of any type of data. Hence our ML model can efficiently handle the dataset and draw concrete conclusions.

With the help of this technique, organizations can predict accurate insights, which can boost their business and seal the deals.

5. REFERENCE

- Shona McCombes. (2021) Sampling Methods | Types, Techniques & Examples *Blog*. [online] Available at: https://www.scribbr.com/methodology/sampling-methods/ [Accessed 15 Dec. 2022].
- Mohamed Elfil and Ahmed Negida (2021) Sampling methods in Clinical Research; an Educational Review https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810526/#:~:text=Sampling%20efficiency%20is%20the%20amount,a%20stratum%20and%20result%20in [Accessed 16 Dec. 2022].
- Prakhar Mishra. (2021) 8 Types of Sampling Techniques *Blog*. [online] Available at: 8 Types of Sampling Techniques. Understanding Sampling Methods by Prakhar Mishra | Towards Data Science [Accessed 17 Dec. 2022]
- CueMath. Methods of Sampling *Blog*. [online] Available at: Methods of Sampling - Types, Techniques, Examples (cuemath.com) [Accessed 18 Dec. 2022].