2025-06-05

Dealing with large data sets: the Data Nugget Subset Selection approach

Vipin Kumar
University of Cassino and Southern Lazio, vipin.kumar@studentmail.unicas.it

Simona Balzano
University of Cassino and Southern Lazio, s.balzano@unicas.it

Giovanni C. Porzio
University of Cassino and Southern Lazio, porzio@unicas.it

Dealing with large data sets: the Data Nugget Subset Selection approach

Vipin Kumar*
vipin.kumar@studentmail.unicas.it
University of Cassino
and Southern Lazio
European University of Technology
(EUt+)
Cassino, Italy

Simona Balzano
s.balzano@unicas.it
University of Cassino
and Southern Lazio
European University of Technology
(EUt+)
Cassino, Italy

Giovanni C. Porzio
porzio@unicas.it
University of Cassino
and Southern Lazio
European University of Technology
(EUt+)
Cassino, Italy

Abstract

Analysing big data has always been a major issue because its massive volume poses significant challenges for traditional analytical techniques. When the number of instances is extremely large, existing approaches become computationally infeasible due to the complexity of many algorithms, along with memory and time constraints inherent in processing large datasets. In such cases, using a subset of the data is considered a more practical solution, and analyses are typically performed over a simple random sample drawn from the entire dataset. Various subsampling methods have been proposed to address these issues. However, they often fall short in producing representative subsamples that can be used across different analytical techniques. Within this framework, this work presents a novel approach based on the so-called Data Nuggets to obtain a subset that can be used for any further required analysis. Existing techniques to get a subset from a large dataset may focus on particular objectives, or they may struggle to capture the structure and statistical properties of the original data. Our method builds on the robustness of the Data Nugget approach, which summarizes a dataset while keeping its structure. This new method, which we name DN-subset selection, is based on sampling from each refined data nugget, finally yielding a much smaller dataset that well represents the whole original large sample. The effectiveness of our method is evaluated through a simulation study.

Keywords

Classification, Data Nuggets, Sub-Sampling

1 Data reduction and Data Nuggets: the DN-Subset Selection approach

The challenge of analysing big data, characterized by a large number of observations (N), is a significant issue for many data analysis methods. Just as an example, traditional clustering methods like K-means and hierarchical clustering become impractical because their computational complexity scales as $P \times N(N-1)/2$. In fact, these methods were developed before the era of big data and cannot handle the massive datasets generated today.

As big data is used extensively in a variety of fields, such as marketing, finance, medical research, and many others, it is essential to rely on some data reduction methods. To address these kinds of challenges, a very common approach is to select a subset of size $n << N$ of the original data set.

Many studies have focused on choosing a statistically significant subset that preserves the distributional and statistical characteristics of the entire dataset. Among many, Chaudhuri et al. (1994) [2] introduced two algorithms to find the best representative points, called seed points. However, these methods rely on the heuristic choice of a specific parameter (called w). On the other hand, if the number of representative points is fixed beforehand, there is no guarantee of obtaining a satisfactory result.

Later, Daszykowski et al. (2002) [5] provided an overview of subset selection methods and introduced a new subset concept using the K-means clustering to define representative subsets. Subsequently, Chen et al. (2007) [3] proposed a method called mIPW-SSWD-PLS, which combines modified iterative predictor weighting for variable selection with a sample set weighted distance for selecting representative samples. Meanwhile, Feldman et al. (2013) [6] made significant theoretical contributions by introducing the concept of constant-size coresets for K-means clustering, PCA, and projective clustering. They showed that these coresets could be effectively used in streaming algorithms.

However, these methods are heavily task-specific designed and highly parameter sensitive. This limits their adaptability and effectiveness across diverse datasets and analysis goals.

The paper by Salloum et al. (2019) [7] introduces the Random Sample Partition (RSP) distributed data model, which divides original data into disjoint blocks that preserve the statistical properties of the entire dataset and that are used as representative sample to perform the analysis. They demonstrated the statistical and computational advantages of the RSP blocks.

Finally, to overcome the weaknesses of existing methods, Cherasia et al. (2022) [4] and Beavers et al. (2025) [1] introduced the Data Nuggets technique, which reduces huge datasets maintaining the data structure and retaining key characteristics, such as the data nuggets' centre, weight, and scale, thereby preserving important information from the original dataset. Data Nuggets provide centre points, which are sets of values computed on the data. In other words, they are not a subsample but syntheses of the data, then can assume values not present in the dataset.

Both the Random Sample Partition and the more recent Data Nugget approach, along with many other methods, work on a set of representative values (e.g., the centers of a set of clusters, or the data nuggets centers themselves). However, it may be convenient to use subsets of the original data as subsamples preserve better the original data distribution. For example, they may include outliers and complex patterns, making such an approach more suitable for the analysis of non-linear structures. Furthermore, subsamples facilitate studies that cannot be performed over some representative values (e.g., bootstrapping procedures).

1.1 The proposal: DN-subset selection

We introduced a novel approach to obtain a subset from a large sample, based on Data Nuggets method. This method reduces the size of big data into smaller collection of representative values based on a chosen distance metric.

Each data nugget summarizes a subset of original data and is characterized by data nugget's centre, weight, and scale. They are designed to maintain the original data structure, which traditional techniques may fail to preserve, especially at the edges. In fact, once a first raw set of nuggets is obtained, a refining analysis makes them as spherical as possible and minimizes their within variability. This is achieved by splitting the nuggets based on an eigenvalue criterion. But, as we discussed above, the Data Nuggets represent manageable collections of points rather than actual subsets.

Our novel approach DN-subset selection addresses this limitation. To get the representative subset, we used the proportional stratified sampling method based on refined data nugget. In this approach, each data nugget is treated as a separate stratum, and a random sample is taken from each stratum based on its proportion in the overall population. The selection is made in a way such that at least one point comes from each nugget (it may happen a nugget is made by a single or by very few points).

Our technique enables the creation of a subset that preserves the distributional and statistical characteristics of the original data, making it suitable for any kind of statistical analysis. The proposed procedure yields a subset of the original points along with a weight for each point corresponding to the inverse of the inclusion probability. Any subsequent analysis must take into account such weights.

To illustrate the effectiveness of our proposal, a small simulation study was conducted in analogy with the work by Cherasia et al. (2022) [4]. Compared to simple random sampling, our method delivers more accurate and consistent parameter estimates across all the simulation settings. This was especially evident when the number of features increased, showing the method's robustness and scalability. In summary, our DN-subset selection strategy offers a reliable and efficient alternative that outperforms simple random sampling, even in high-dimensional settings. Future work could explore extending the DN-subset selection method to real-world datasets to assess its practical effectiveness across different analytical method.

References

[1] Traymon E Beavers, Ge Cheng, Yajie Duan, Javier Cabrera, Mariusz Lubomirski, Dhammika Amaratunga, and Jeffrey E Teigler. 2025. Data nuggets: A method for reducing big data while preserving data structure. *Journal of Computational and Graphical Statistics* 34, 1 (2025), 330–342.

[2] Debasis Chaudhuri, CA Murthy, and BB Chaudhuri. 1994. Finding a subset of representative points in a data set. *IEEE transactions on systems, man, and cybernetics* 24, 9 (1994), 1416–1424.

[3] Da Chen, Wensheng Cai, and Xueguang Shao. 2007. Representative subset selection in modified iterative predictor weighting (mIPW)—PLS models for parsimonious multivariate calibration. *Chemometrics and intelligent laboratory systems* 87, 2 (2007), 312–318.

[4] Kenneth Edward Cherasia, Javier Cabrera, Luisa T Fernholz, and Robert Fernholz. 2022. Data nuggets in supervised learning. In *Robust and Multivariate Statistical Methods: Festschrift in Honor of David E. Tyler*. Springer, 429–449.

[5] Michal Daszykowski, Beata Walczak, and DL Massart. 2002. Representative subset selection. *Analytica chimica acta* 468, 1 (2002), 91–103.

[6] Dan Feldman, Melanie Schmidt, and Christian Sohler. 2020. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.* 49, 3 (2020), 601–657.

[7] Salman Salloum, Joshua Zhexue Huang, and Yulin He. 2019. Random sample partition: a distributed data model for big data analysis. *IEEE Transactions on Industrial Informatics* 15, 11 (2019), 5846–5854.