Adaptive Granulation: Data Reduction at the Database Level

Hossein Haeri[1][a], Niket Kathiriya[2][b], Cindy Chen[2][c] and Kshitij Jerath[1][d]
[1]*Department of Mechanical Engineering, University of Massachusetts Lowell, Lowell, MA*
[2]*Department of Computer Science, University of Massachusetts Lowell, Lowell, MA*
{*hossein_haeri, niket_kathiriya, cindy_chen, kshitij_jerath*}*@uml.edu*

Keywords: Data Granulation, Data Reduction, Data Aggregation, Training Set Size Reduction

Abstract: In an era where data volume is growing exponentially, effective data management techniques are more crucial than ever. Traditional methods typically manage the size of large datasets by reducing or aggregating data using a pre-specified granularity. However, these methods often face challenges in retaining vital information when dealing with large and complex datasets, especially when such datasets reside in databases. We propose a novel and innovative approach called Adaptive Granulation that addresses this issue by performing data reduction or aggregation at the database level itself. A key concern that arises in the data reduction process is the potential trade-off between the reduction of data volume and the preservation of prediction accuracy. This is particularly relevant in scenarios where the primary goal is to leverage the reduced dataset for predictive modeling. Our method employs Allan variance, originally developed for frequency stability analysis of atomic clocks, to dynamically adjust the granularity of data aggregation based on the inherent structure and characteristics of the dataset. By minimizing bias across different scales, Adaptive Granulation effectively manages trade-offs between diverse aspects of the data such as underlying patterns, noise levels, and sampling density. This paper outlines the algorithmic strategies for implementing Adaptive Granulation at the database level and assesses its performance through the reduction of the training set size for a downstream regression task on a variety of real-world and synthetic datasets. The results indicate that our method can adaptively optimize granule sizes to effectively balance data patterns, noise levels, and sample densities across the entire data space. Adaptive Granulation thus represents a significant advancement for efficient data management and reduction in the big data era.

1 Introduction

Data reduction methods are increasingly becoming vital in our data-driven world. As the proliferation of big data continues across various sectors, including healthcare (Raghupathi and Raghupathi, 2014), finance (Hasan et al., 2020), social media (Sahatiya, 2018), e-commerce (Akter and Wamba, 2016), traffic (Lv et al., 2015; Kim and Jerath, 2022) and more (Chen et al., 2014), the sheer volume of data we generate is staggering. For example, in healthcare, patient records, wearable device outputs, and genomic data produce immense amounts of data. Social media platforms like Facebook and Twitter generate millions of posts and interactions daily. Retail giants like Amazon collect data about customer behavior, preferences, and purchasing habits on a massive scale. It is estimated that we are generating more than 2.5 quintillion bytes of data each day (Zicari, 2012), and this rate is only expected to increase with advancements in technology and increased internet accessibility. By 2025, it is projected that there will be more than 180 zettabytes of data in the world (Holst, 2021).

Utilizing such large amounts of data for inference, as well as storing them in databases, is increasingly becoming an untenable exercise. While big data can provide valuable insights and aid decision-making, it can also overwhelm our computational and memory resources, thus hindering effective analysis. While studies have explored the use of cloud-based and edge-based units, training data-driven models on extensive datasets presents multiple challenges. Firstly, we require significant computational resources, including substantial processing power and memory. This potentially constrains usage by individuals or organizations with limited computational capabilities.

[a] https://orcid.org/0000-0002-6772-6266
[b] https://orcid.org/0000-0002-4146-3402
[c] https://orcid.org/0000-0002-8712-8108
[d] https://orcid.org/0000-0001-6356-9438

Secondly, the time to train a model can substantially increase with larger datasets, causing potential delays in the implementation and deployment of models that may be critical for time-sensitive applications. Finally, while more data generally reduces overfitting, if the dataset contains a high degree of noise or irrelevant features, models could potentially overlearn these aspects, negatively impacting performance on unseen data.

Therefore, data reduction methods, which seek to reduce the volume of data while preserving its critical information, are becoming increasingly important. These methods allow us to manage and analyze large datasets more efficiently and effectively, thereby unlocking the potential of big data while avoiding the pitfalls of data overload. Then, data reduction is not just a practical necessity; it is a critical step in transforming raw data into meaningful insights.

To address these complexities induced by big data, we present a novel and innovative approach to data reduction at the database level itself, which we term as **Adaptive Granulation (AG)**. Our approach provides a systematic way to balance the competing desires to maintain high data fidelity (and hence increase model predictive power) and to reduce the size of the data (and hence speed up training times). Our novel approach is inspired by the concept of Allan variance, which was originally developed to manage similar competing interests while examining the frequency stability of atomic clocks (Allan, 1966). As an additional advantage, this approach enables the reduction of the database size as well, resulting in faster approximate query response times. Thus, Adaptive Granulation excels in the context of big data by offering an efficient and effective means to manage and process large volumes of data for training data-driven models, while leveraging the capabilities of the database itself.

2 Related Work

Data reduction, especially in large datasets, is crucial for improving the efficiency of data-driven models (Sandhu, 2021). Classical methods include simple random sampling, stratified sampling (Zhang et al., 2022), and reservoir sampling (Kim et al., 2020). However, with the rise of big data, methods like record-level and block-level sampling have gained prominence (Hasanin et al., 2019). The former can be less efficient on distributed data, while the latter, using traditional partitioning, may not always provide representative samples (Mahmud et al., 2020). In this work, we introduce a new block-level sampling technique optimized for enhanced performance in regression tasks.

Researchers in the field of machine learning have primarily approached this methodology from perspectives that are specific to the machine learning algorithms themselves, i.e., there exist data reduction methods for specific applications to Support Vector Machines (SVMs), decision trees, and neural networks, to name a few (Alwajidi, 2020; Mahmud et al., 2020; ur Rehman et al., 2016). Here, we briefly review some of these methods from the viewpoint of the ML algorithms.

Support Vector Machines: For SVMs, one of the main challenges is to reduce the number of support vectors, which determine the decision boundary and affect the computational cost and generalization ability of the model. Several approaches have been proposed to select a subset of support vectors or training samples that can approximate the original decision boundary with minimal loss of accuracy (Birzhandi et al., 2022). For example, prior works have demonstrated the use of clustering-based techniques to identify and remove non-relevant samples that are far from the decision boundary (Koggalage and Halgamuge, 2004; Yao et al., 2013; Santana et al., 2020). Ghaffari has proposed a method to divide the training set into boundary, non-boundary, and harmful patterns, and then select representatives of non-boundary data and combine them with boundary patterns to form a reduced set (Ghaffari, 2021). Other methods include using information entropy (Zhan and Shen, 2005), chunking (Hsieh et al., 2008), or regression (Osuna and Girosi, 1998) to reduce the number of training samples for SVMs. Most of these data reduction methods rely on a fundamental characteristic of the SVM: that learning with the SVM algorithm is dependent on only a few support vectors, as compared to the totality of available data.

Decision Trees: For decision trees, one of the main challenges is to manage the tree size and complexity, which affect the interpretability and generalization ability of the model. Several approaches have been proposed to prune or simplify the tree structure after or during the tree construction process. For example, (Oates and Jensen, 1997) studied the effects of training set size on decision tree complexity and showed that increasing training set size often results in a linear increase in tree size, even when that additional complexity results in no significant increase in classification accuracy. They argued that random data reduction is a baseline against which more sophisticated data reduction techniques should be compared. Other methods include using misclassification costs (Bradford et al., 1998), Laplace correction (Brodley and Friedl, 1999), or error-based pruning (Peng et al.,

2021) to prune decision trees.

Neural Networks: For neural networks, one of the main challenges is to handle large-scale vision-language pre-training tasks, which require huge amounts of data and computational resources. Several approaches have been proposed to reduce the data size or complexity for such tasks. For example, (Jinpeng Wang et al., 2023) proposed a method called Too Large; Data Reduction for Vision-Language Pre-Training (TL;DR), which uses a two-stage process to select a subset of image-text pairs that are informative and diverse for pre-training vision-language models. They showed that their method can achieve comparable or better performance than existing methods with much fewer data and computation time.

Adaptive Granulation differs from the existing data reduction methods in several ways. **First**, it is a systematic preprocessing method that can be utilized by any data-driven model, such as SVMs, decision trees, or neural networks. Unlike the methods that are tailored for specific models or tasks, Adaptive Granulation does not rely on any model-specific assumptions or parameters. **Second**, it is a flexible method that can dynamically adjust the granularity level based on the underlying structure and characteristics of the data using Allan variance. Unlike the methods that use a fixed or uniform level of granularity across the entire dataset, Adaptive Granulation can accommodate variations in pattern complexity, noise, and sampling density within different sections of the dataset. **Finally**, our method has the additional advantage that it fundamentally operates in and leverages the database itself. This leads to a significant simplification of data reduction, storage, and retrieval: all precursors to subsequent machine learning steps. These differences make Adaptive Granulation a powerful and versatile data reduction method that can effectively reduce data volume while preserving crucial information inherent in the dataset. In the next section, we discuss the method in more detail.

3 Adaptive Granulation

Adaptive Granulation is an innovative approach to data reduction that aims to effectively reduce data volume while preserving crucial information inherent in the dataset. This method is particularly useful in the context of big data, where handling and processing large amounts of data efficiently and effectively becomes challenging.

By dynamically adjusting the granularity level, Adaptive Granulation aggregates data points based on their underlying structure and characteristics (Maddipatla et al., 2023; Maddipatla et al., 2021). The *granularity* here refers to the level of detail or scale at which the data is considered or analyzed. For instance, in a one-dimensional temporal dataset, granularity might refer to whether the data is examined on a yearly, monthly, daily, or hourly basis. It is important to note that the algorithm automatically determines the optimal level of granulation (Sinanaj et al., 2022; Haeri et al., 2021; Haeri et al., 2022).

The overarching workflow of this study is depicted in Figure 1 which can be summarized as the following. Initially, a vast assortment of data points is incorporated as raw data into the database. These data points subsequently serve as the foundation for the creation of an R* tree, a spatial indexing mechanism that facilitates hierarchical granulation or clustering of data into nested minimum bounding rectangles (MBRs). We then introduce Allan variance as an innovative method to adaptively granulate the data by pinpointing the optimal scale of data granularity. This is achieved by identifying the level of granules that exhibits the minimum Allan variance in comparison to their parent and child granules, as will be discussed in Algorithm 1. The outcome of this procedure is a series of non-nested granules that constitute a reduced dataset. In the final step of our approach, the reduced dataset is formed by computing the centroid as well as the average value of the target attribute. This transformation allows us to retain the most descriptive characteristics of the data (i.e., information that can be effectively used for a downstream prediction task) while significantly reducing the volume, aiding in the efficient analysis and modeling of large datasets.

3.1 R* tree as a Hierarchical Clustering Data Structure

Adaptive Granulation leverages the R* tree structure used to organize multi-dimensional large data sets. First, data is inserted into an R* tree. This process involves splitting and adjusting nodes in the tree to accommodate the new data while maintaining the tree's balance and minimizing the overlapping area of the MBRs. More details about the process of R* tree creation can be found in (Beckmann et al., 1990). The R* tree structures data points into a nested set of Minimum Bounding Rectangles (MBR) where each MBR represents a region in the multi-dimensional space that contains one or more data points or other MBRs. In the following discussion, we assume that $\mathbf{x} \in \mathcal{X}$ represents an input vector in a d_x-dimensional feature space, and $\mathbf{y} \in \mathcal{Y}$ represent the target attribute vector in d_y-dimensional space, which we are attempting to predict. Now, we begin with the large dataset

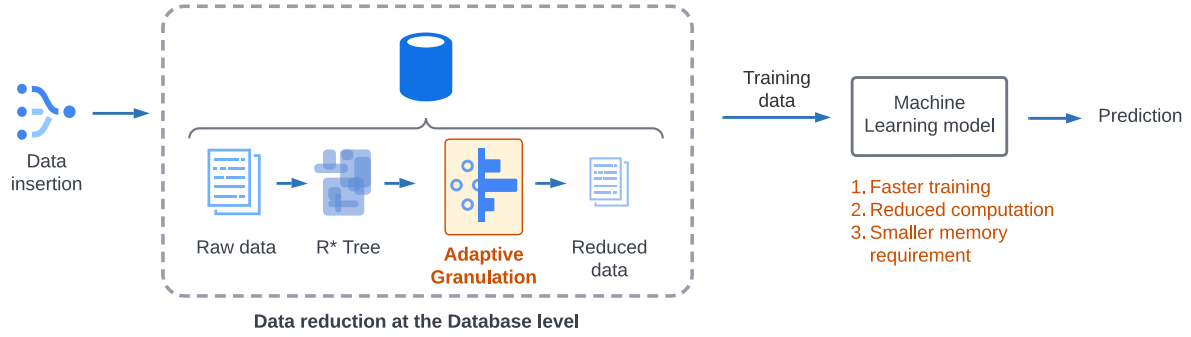

Figure 1: Adaptive Granulation can be deployed at the database level in order to reduce the size of training data set.

$\mathcal{D} = \{(\mathbf{x_i}, \mathbf{y}_i)\}$ where $i = \{1, 2, \ldots, N\}$ and N is the total number of data points in the original dataset. Next, we seek to use the Adaptive Granulation algorithm to create a representative reduced data set $\mathcal{D}_{\text{red}} = \{(\bar{\mathbf{x}}_j, \bar{\mathbf{y}}_j)\}$ containing aggregated data points, where $j = \{1, 2, \ldots, N_{\text{red}}\}$ and N_{red} denotes the total number of data points in the reduced data set, with the expectation that the AG algorithm will systematically lead to $N_{\text{red}} \ll N$. We define the *reduction ratio* as $\zeta = 1 - N_{\text{red}}/N$, to represent the magnitude of data that was removed from the original dataset.

Let $\mathcal{M}$ be the set of all the MBRs in the R* tree. Each MBR, $m \in \mathcal{M}$, is identified by a tuple $(r_m, \mathbf{x}_m^{\min}, \mathbf{x}_m^{\max}, C_m, p_m, \bar{y}_m, \sigma_m^2)$, where

- $r_m \in \{0, 1, 2, \ldots\}$ denotes the level of the MBR in the R* tree. In our work, level $r = 0$ represents the actual data points, and the level number increases as we move from the tree leaves towards the root.
- $\mathbf{x}_m^{\min}$ and $\mathbf{x}_m^{\max}$ are d_x-dimensional vectors representing the minimum and maximum bounds of the MBR in the multi-dimensional feature space,
- $C_m \subset \mathcal{M}$ is the set containing all the children of MBR m, and the children themselves are MBRs in the R* tree residing at level $r_m - 1$,
- $p_m \in \mathcal{M}$ represents the parent of the MBR m, and the parent is also an MBR in the R* tree residing at level $r_m + 1$,
- $\bar{\mathbf{y}}_m$ is a d_y-dimensional vector representing the average of the target attribute vectors of the children of the MBR m,
- $\sigma_m^2 \in \mathbb{R}^+$ is the Allan variance of the MBR m, and is a measure of the variability within the MBR which is discussed in the next subsection.

3.2 Using Allan variance to Determine the Granularity Level

The notion of Allan variance (AVAR) was first introduced by David Allan as a way to study the statistics of frequency stability in precision oscillators such as atomic clocks (Allan, 1966). The fundamental idea seeks to address the phenomenon where the oscillatory behavior of such systems can be noisy as well as drift at the same time. The competition between noise and signal makes it difficult to determine the optimal timescale that should be used to ascertain the frequency of the atomic clock. In the context of the current work, this can be thought of as being similar to balancing two aspects: (a) minimizing the effects of noise in the data reduction process, while (b) maximizing the ability to capture the underlying patterns (or 'drift') present in the data, even as we reduce its size. For example, if data reduction is performed via aggregation at too small of a scale, then a significant amount of noise will remain in the reduced data set, representing a missed opportunity to further reduce the size of the dataset (Yang et al., 2021). Similarly, if data reduction is performed via aggregation at too large of a scale, we will manage to significantly reduce the size of the data, but at the cost of poorer predictive performance. In such scenarios, Allan variance can help us determine the optimal scale that minimizes the bias in the averaged signal (Jerath et al., 2018; Haeri et al., 2021; Sinanaj, 2021; Haeri et al., 2022), in effect providing a data reduction or aggregation scale that optimally groups data points.

Thus, leveraging the principles of Allan variance, we can propose a measure of stability across a dataset's feature space. Prior studies have corroborated the efficacy of Allan variance in discerning the ideal granule size for temporal database granulation (Sinanaj et al., 2022; Sinanaj, 2021). Further, it has been demonstrated that employing this approach facilitates optimal averaging scales for moving average estimation tasks (Haeri et al., 2022; Haeri et al., 2020). In the present work, we build on these foundations, charting new territory in two main directions. As an important distinction from other previous works, we extend the concept of granulation to a

multi-dimensional data environment. Equally as importantly, we introduce an adaptive mechanism to determine the granularity scale, enabling dynamic data reduction in heterogeneous characteristics observed across the feature space of the dataset.

The Allan variance for the MBR m can be calculated as

$$\sigma_m^2 = \frac{1}{|C_m|} \sum_{c \in C_m} ||\bar{\mathbf{y}}_m - \bar{\mathbf{y}}_c||^2 \tag{1}$$

where c is a child MBR in the set C_m, and $\bar{\mathbf{y}}_c$ is its average target attribute. Further, for each MBR m, $\bar{\mathbf{y}}_m$ is obtained by

$$\bar{\mathbf{y}}_m = \frac{1}{|C_m|} \sum_{c \in C_m} \bar{\mathbf{y}}_c. \tag{2}$$

.

In this framework, Allan variance serves as a metric to assess the variability in multi-dimensional data specific to the level of an MBR. It is important to highlight that, in our research, σ_m^2 signifies the variance of the average values from the child MBRs, rather than the variance of individual data points within that MBR. This variance provides insight into the dispersion of the data points of child MBRs nested within the MBR m. Depending on the characteristics of the dataset, these calculations may produce different values of the Allan variance (a) for MBRs at different levels r in the R* tree, but in the same general region of the feature space, and (b) for MBRs at the same level r, but in different regions of the feature space. A large value of Allan variance in a specific MBR at a specific level can be indicative of one or more of the following phenomena:

(a) Rapid change in underlying data pattern. If the data pattern is changing rapidly at a specific scale and feature space region, the AVAR of the associated MBR may be large. Using aggregated data of this MBR as the reduced data point will lead to a failure to capture the underlying pattern in the dataset. This will produce smaller-sized, but lower fidelity reduced datasets.

(b) Significant presence of noise. If the data is very noisy at a specific scale and feature space region, then the AVAR of the associated MBR may also be large. Using aggregation to reduce data in this situation will lead to capture of lot of noise but less meaningful data patterns. This will produce higher fidelity, but larger datasets without any meaningful data reduction.

By opting for the level of granulation linked with the minimum Allan variance, this trade-off is effectively managed – striking a balance between data reduction and preservation of essential information. We discuss this further in Section 4. The overall procedure for implementing Adaptive Granulation is depicted in Algorithm 1. The algorithm begins by requiring a list $\mathcal{M}$ of all the MBRs in the tree which are sorted in ascending order according to their level, r_m. In the first loop the Allan variance of each MBR is calculated based on Equation (1) which itself requires calculating averages using Equation (2). If the MBR is at level 0 (indicating it is a leaf node), the granulation flag is set to `True`; otherwise, it is set to `False`. This initialization prepares the MBRs for the subsequent steps of the algorithm. In the subsequent loop, for each MBR in $\mathcal{M}$, if every child of the current MBR has its granulation flag marked as `True`, and the Allan variance of this MBR is smaller than that of any of its children, the current MBR's granulation flag is updated to `True`, while the flags of all its children are reset to `False`. This step ensures the MBR that is selected as a granule has a variance (as defined in Equation 1) that is smaller than both its parent and any of its children, and thus represents the optimal granulation level. This process can also be thought of as elevating the granulation flag to a higher level, suggesting that the target attribute exhibits less variability at a broader granulation scale. The algorithm then concludes by returning all MBRs whose granulation flag is True. These selected MBRs represent the data points that have been adaptively granulated, and they form the reduced dataset for further analysis or modeling tasks. According to Algorithm 1, each data point in the R* tree will be encompassed by a singular MBR with the True granulation flag. This holds as every data point starts with a true flag, and when the algorithm opts to transfer this flag to the parent-level MBR, all its child nodes relinquish their flags. Such systematically-generated granules will provide the best trade-off between data reduction (eliminating noise) and data retention (keeping data patterns), irrespective of the eventual prediction algorithm that will be used.

4 Results and Discussion

In this section, the performance of Adaptive Granulation as a data reduction method is evaluated for three synthetic and six real-world datasets. We benchmark the efficacy of Adaptive Granulation against random sampling—a fundamental baseline in data reduction that, despite its simplicity, has been demonstrated to outperform many advanced sampling methods in the literature (Oates and Jensen, 1997; Hasanin et al.,

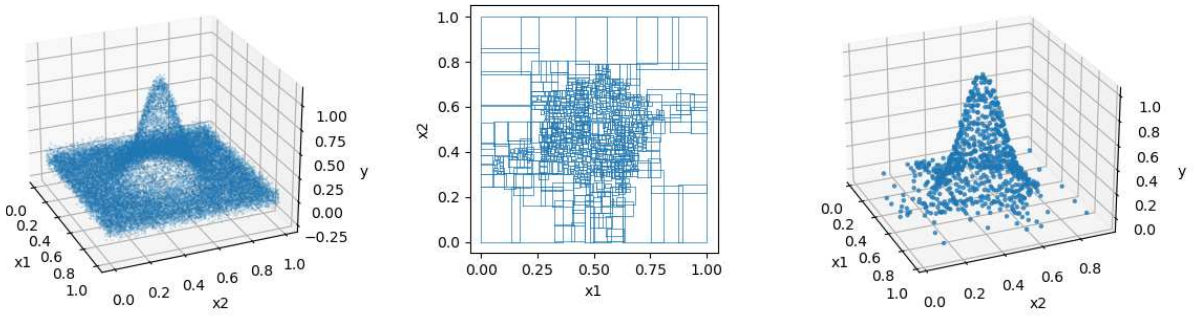

Figure 2: A synthetic data aggregation scenario with varying underlying pattern complexity. (a) Input data points. (b) Generated granules using Adaptive Granulation. (c) Granule centroids (y value denotes average granule target attribute).

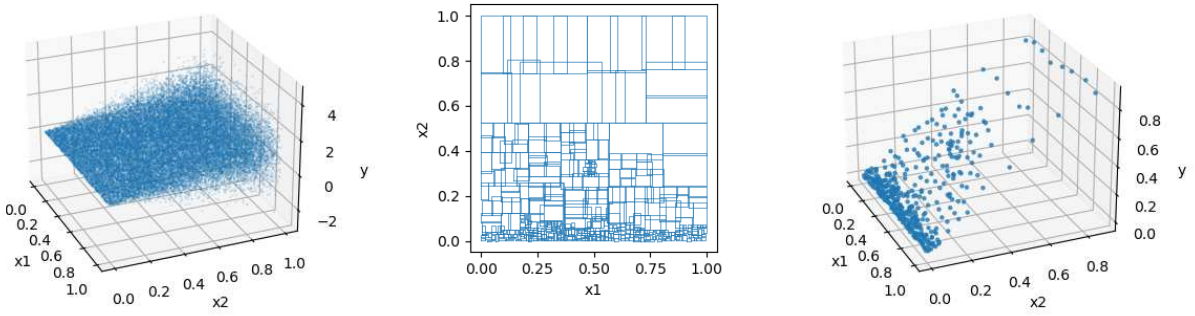

Figure 3: A synthetic data aggregation scenario with varying noise characteristics along the x_2 axis. (a) Input data points. (b) Generated granules using Adaptive Granulation. (c) Granule centroids (y value denotes average granule target attribute).

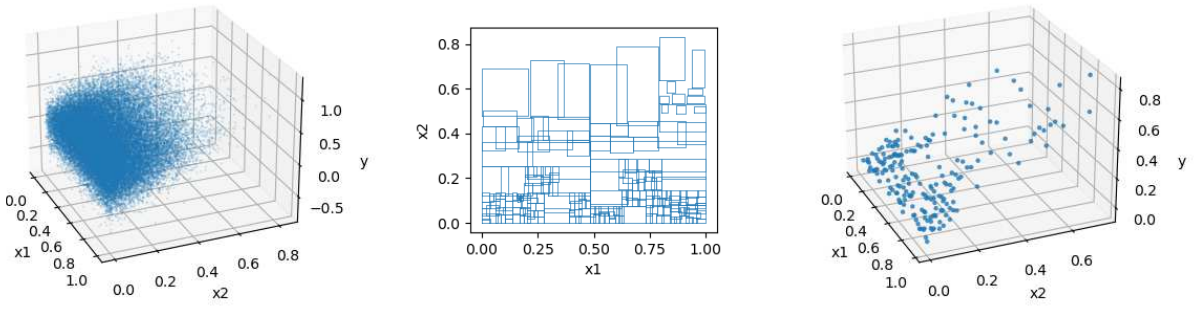

Figure 4: A synthetic data aggregation scenario with varying data density along the x_2 axis. (a) Input data points. (b) Generated granules using Adaptive Granulation. (c) Granule centroids (y value denotes average granule target attribute).

Algorithm 1 Adaptive Granulation

Require: List of all MBRs, $\mathcal{M}$, in the R* tree
for each ordered m in $\mathcal{M}$ **do**
  Calculate σ_m^2 according to Eq. (1) and (2)
end for
Set m.`is_granule` to `True` for MBRs at level 0 and to `False` for all others.
for each ordered m in $\mathcal{M}$ **do**
  if all children c in C_m: are granules, and have $\sigma_m^2 < \sigma_c^2$ **then**
    m.`is_granule` $\leftarrow$ `True`
    for each c in C_m **do**
      c.`is_granule` $\leftarrow$ `False`
    end for
  end if
end for
return MBRs in $\mathcal{M}$ with `is_granule` == `True`

2019).

Our experimental setup involves comparing the Mean Absolute Error (MAE) of various machine learning models. The complete dataset is initially split into train and test datasets, with the test data constituting 30% of the total. The Adaptive Granulation method is then applied to the training data, using various fan-out values for the R* tree, which represents the maximum number of children within a single MBR. The centroids of the resulting granules serve as data points in the reduced dataset. As a baseline comparison, an equivalent number of data points are sampled randomly from the training data. The evaluation process is performed 100 times for each fan-out value, which varies from 8 to 64 with a step size of 4. Model tuning is performed using the `RandomSearchCV` function from the `sklearn` library, with 100 iterations and 3-fold cross-validation.

4.1 Synthetic Scenarios

Unlike conventional data reduction techniques that apply a fixed level of granularity across the entire dataset, Adaptive Granulation is flexible. It recognizes that different portions of the dataset might require different levels of granularity depending on their underlying pattern complexity, noise, and density. In this subsection, we assess the efficacy of Adaptive Granulation through three synthetic datasets. For these, data points are synthesized from the joint distribution $f(x_1, x_2, y)$ where x_1 and x_2 are independent features, drawn from distributions f_1 and f_2, respectively. The target attribute, y, is produced in accordance with a deterministic pattern function, supplemented by a stochastic noise function. In the subsequent sections, we outline the setup for these scenarios and explore the influence of these three factors on the determination of granule sizes.

1. **Underlying pattern complexity:** When the complexity of the patterns in the data is high, small granules (i.e., a finer granularity) may be needed to capture the details of these patterns. High complexity could be due to non-linear relationships between variables, the presence of many interacting variables, or rapidly changing patterns. In such cases, a coarse granularity could oversimplify the data and fail to capture these complexities, resulting in the loss of crucial information. In this experimental scenario, f_1 and f_2 are independent uniform distributions ($U(0,1)$) and the noise function is a zero mean Gaussian distribution which is consistent across both x_1 and x_2 dimensions ($\mathcal{N}(\mathbf{0}, 0.05\mathbf{I})$). The target values y are generated according to a squared exponential Gaussian pattern function located at $(0.5, 0.5)$. As shown in Figure 2, Adaptive Granulation selects smaller granules in regions where the Gaussian dome is rising, capturing the details of the pattern. Conversely, in areas where the underlying data pattern exhibits less variation, the Adaptive Granulation process opts for larger granule sizes, effectively balancing the granularity level with the inherent complexity of the data.
2. **Noise:** Noise refers to random or irrelevant variation in the data. When the level of noise in the dataset is high, it can be beneficial to use larger granules (i.e., a coarser granularity). This is because a finer granularity could risk overfitting to the noise, leading to less accurate or less meaningful results. By aggregating the data at a coarser level, Adaptive Granulation can smooth out the noise and reveal the underlying patterns or trends in the data. Conversely, when the noise level is low, a finer granularity may be used to capture more detailed patterns. In this scenario, f_1 and f_2 are independent uniform distributions ($U(0,1)$). The target attribute, y, is determined by a linear pattern function across the x_2 dimension (i.e., $y = x_2$) and is supplemented by zero mean Gaussian noise. This noise remains consistent across the x_1 dimension but varies along the x_2 dimension, specifically conforming to $\mathcal{N}(0, x_2)$. As illustrated in Figure 3, Adaptive Granulation tends towards the selection of larger granules as the magnitude of noise escalates.
3. **Sampling Density:** Sampling density refers to the number of data points in a given space. When the sampling density is high, there are many closely packed data points, and a finer granularity may be

needed to capture the variation within this dense space. Conversely, when the sampling density is low, the data points are sparse and spread out, and a coarser granularity might suffice to capture the essential characteristics of the data. By adjusting the granule size based on the sampling density, Adaptive Granulation ensures that neither the dense nor the sparse areas of the dataset are under or over-represented. In this scenario, f_1 is a uniform distribution ($U(0,1)$), however, f_2 is a half-normal distribution (i.e., $|\mathcal{N}(0,0.2\,x_2)|$). The target attribute, y, is determined by a linear pattern function across the x_2 dimension (i.e., $y = x_2$) and is supplemented by zero mean Gaussian noise which is consistent across both x_1 and x_2 dimensions ($\mathcal{N}(0,0.2)$).

4.2 Real-world Scenarios

In this subsection, we have evaluated Adaptive Granulation in six real-world datasets: (1) Abalone, encompassing 4,000 records (Nash et al., 1995); (2) Air Quality, incorporating 9,000 records (Vito, 2016); (3) Bike, containing 10,000 records (Fanaee-T and Gama, 2014); (4) California Housing, which has 20,000 records (Pace and Barry, 1997); (5) Elevators, with 16,000 records; and (6) Metro Interstate Traffic Volume, comprising 48,000 records (Hogue, 2019).

To ensure the integrity of our data prior to analysis, we employ a two-pronged preprocessing strategy. First, we eliminate any outliers using the Interquartile Range (IQR) rule, whereby data points more than 1.5 times the IQR distant from the first or third quartiles are discarded. Second, we perform feature selection based on the importance of each feature. The importance is quantified by the mutual information, calculated using the `sklearn.feature_selection` module. Both preprocessing steps are required since the presence of outliers can distort the resultant granules, whereas the inclusion of non-informative features could lead to the creation of suboptimal or arbitrary granules. From the results of both synthetic data (Figure 5) and real-world data (Figure 6), we conclude that a model trained with Adaptive Granulated data performs better than the same model trained with randomly sampled data.

5 Concluding Remarks

In this study, we have introduced and evaluated Adaptive Granulation, a novel and efficient data reduction technique designed to handle the growing challenges presented by the enormous volumes of data in our digital era. In this work, we have demonstrated that Adaptive Granulation can successfully condense large datasets into dynamically sized granules of aggregated data. The uniqueness of this method lies in its intelligent and systematic aggregation process that takes into consideration the noise characteristics of the dataset, the underlying data patterns, and the data sample density. Importantly, the use of Allan variance in the granulation process helps produce a compact and less demanding dataset while preserving the fidelity of the original information. Perhaps equally as importantly, we emphasize that Adaptive Granulation can be implemented at the database level itself. This feature offers the advantage of managing and reducing data within the storage system, eliminating the need to handle raw, unwieldy datasets directly. This in-situ processing capability makes Adaptive Granulation an attractive solution for efficient data management in data-intensive applications. However, it is important to note that our proposed method is constrained to numerical values, as outlined in this paper, due to the inherent inability of the averaging process to accommodate categorical fields. Our empirical evaluation reveals that the reduced datasets generated through Adaptive Granulation are highly effective for downstream inference tasks. Specifically, our tests have shown that models trained on these refined datasets perform comparably to those trained on randomly sampled datasets, benefiting from a significant decrease in training time and resource requirements. The capacity of Adaptive Granulation to balance the trade-off between data size and pattern preservation, coupled with its ease of implementation within the database, makes it a powerful tool for handling the complexities of Big Data. We expect this approach to enable more efficient utilization of large datasets in training predictive models, ultimately leading to better resource management and improved outcomes in data-intensive domains.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1932138. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

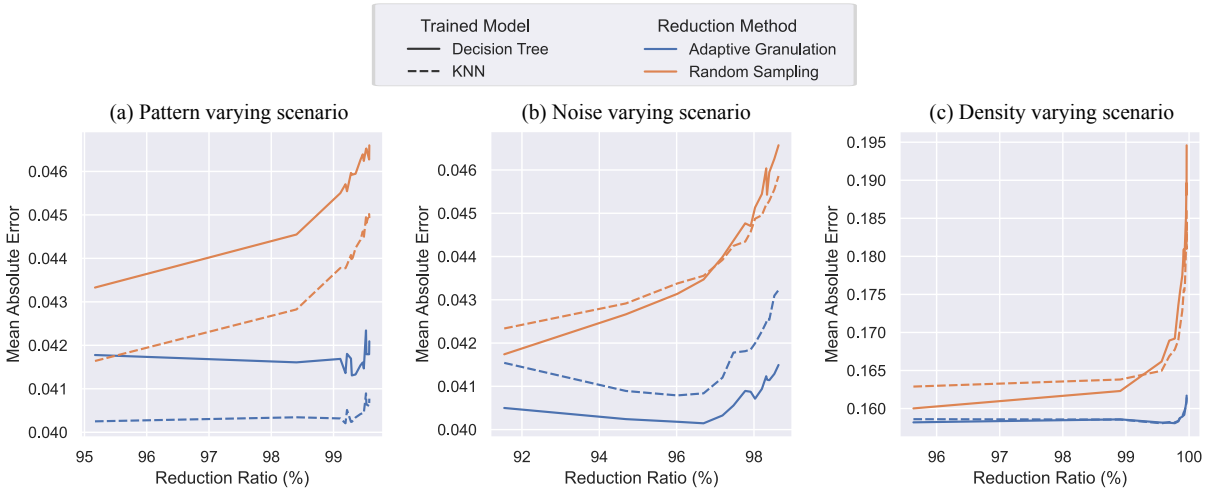

Figure 5: Mean Absolute Error comparison on synthetic datasets. The reported values represent the average results across 100 repeated runs.

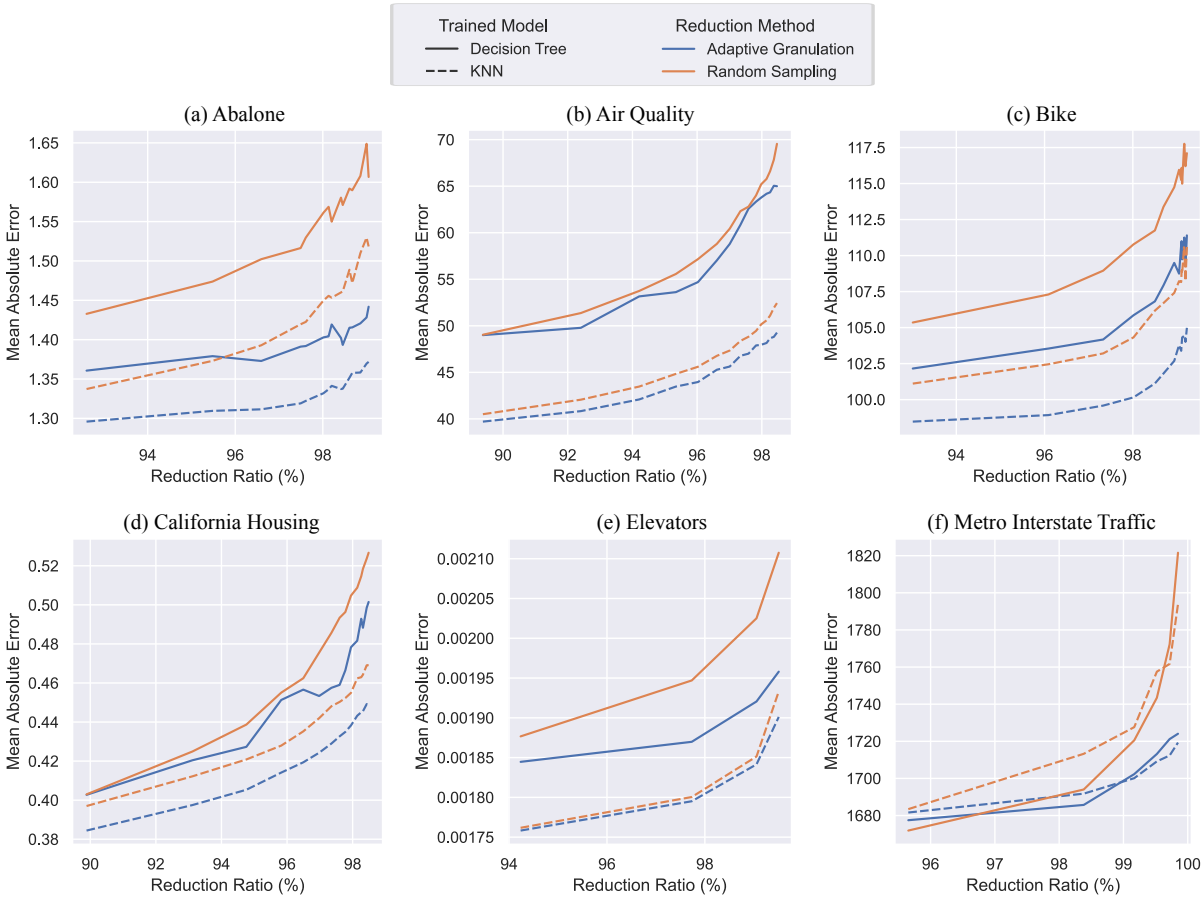

Figure 6: Mean Absolute Error comparison on real-world datasets

REFERENCES

Akter, S. and Wamba, S. F. (2016). Big data analytics in e-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26:173–194.

Allan, D. W. (1966). Statistics of atomic frequency standards. *Proceedings of the IEEE*, 54(2):221–230.

Alwajidi, S. K. (2020). *Hierarchical Aggregation of Multidimensional Data for Efficient Data Mining*. PhD thesis, Western Michigan University.

Beckmann, N., Kriegel, H.-P., Schneider, R., and Seeger, B. (1990). The r*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the 1990 ACM SIGMOD international conference on Management of data*, pages 322–331.

Birzhandi, P., Kim, K. T., and Youn, H. Y. (2022). Reduction of training data for support vector machine: a survey. *Soft Computing*, 26(8):3729–3742.

Bradford, J. P., Kunz, C., Kohavi, R., Brunk, C., and Brodley, C. E. (1998). Pruning decision trees with misclassification costs. In *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings 10*, pages 131–136. Springer.

Brodley, C. E. and Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167.

Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, 19:171–209.

Fanaee-T, H. and Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127.

Ghaffari, H. R. (2021). Speeding up the testing and training time for the support vector machines with minimal effect on the performance. *The Journal of Supercomputing*, 77(10):11390–11409.

Haeri, H., Beal, C. E., and Jerath, K. (2020). Near-optimal moving average estimation at characteristic timescales: An allan variance approach. *IEEE Control Systems Letters*, 5(5):1531–1536.

Haeri, H., Beal, C. E., and Jerath, K. (2021). Near-optimal moving average estimation at characteristic timescales: An allan variance approach. *IEEE Control Systems Letters*, 5(5):1531–1536.

Haeri, H., Soleimani, B., and Jerath, K. (2022). Optimal moving average estimation of noisy random walks using allan variance-informed window length. In *2022 American Control Conference (ACC)*, pages 1646–1651. IEEE.

Hasan, M. M., Popp, J., and Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data*, 7(1):1–17.

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., and Bauder, R. A. (2019). Severely imbalanced big data challenges: investigating data sampling approaches. *Journal of Big Data*, 6(1):1–25.

Hogue, J. (2019). Metro Interstate Traffic Volume. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5X60B.

Holst, A. (2021). Amount of data created, consumed, and stored 2010-2025. *Technology & Telecommunications Retrieved*, pages 06–29.

Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S. S., and Sundararajan, S. (2008). A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th international conference on Machine learning*, pages 408–415.

Jerath, K., Brennan, S., and Lagoa, C. (2018). Bridging the gap between sensor noise modeling and sensor characterization. *Measurement*, 116:350–366.

Jinpeng Wang, A., Qinghong Lin, K., Junhao Zhang, D., Weixian Lei, S., and Shou, M. Z. (2023). Too large; data reduction for vision-language pre-training. *arXiv e-prints*, pages arXiv–2305.

Kim, C. D., Jeong, J., and Kim, G. (2020). Imbalanced continual learning with partitioning reservoir sampling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 411–428. Springer.

Kim, T. and Jerath, K. (2022). Congestion-aware cooperative adaptive cruise control for mitigation of self-organized traffic jams. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6621–6632.

Koggalage, R. and Halgamuge, S. (2004). Reducing the number of training samples for fast support vector machine classification. *Neural Information Processing-Letters and Reviews*, 2(3):57–65.

Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F.-Y. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

Maddipatla, S. P., Haeri, H., Jerath, K., and Brennan, S. (2021). Fast allan variance (favar) and dynamic fast allan variance (d-favar) algorithms for both regularly and irregularly sampled data. *IFAC-PapersOnLine*, 54(20):26–31.

Maddipatla, S. P., Pakala, R., Haeri, H., Chen, C., Jerath, K., and Brennan, S. (2023). Using databases to implement algorithms: Estimation of allan variance using b+-tree data structure. In *Proc. of the Modeling, Estimation, and Control Conf. 2023, Lake Tahoe, NV*.

Mahmud, M. S., Huang, J. Z., Salloum, S., Emara, T. Z., and Sadatdiynov, K. (2020). A survey of data partitioning and sampling methods to support big data analysis. *Big Data Mining and Analytics*, 3(2).

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., and Ford, W. (1995). Abalone. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55C7W.

Oates, T. and Jensen, D. (1997). The effects of training set size on decision tree complexity. In *Sixth International Workshop on Artificial Intelligence and Statistics*, pages 379–390. PMLR.

Osuna, E. and Girosi, F. (1998). Reducing the run-time complexity of support vector machines. In *International Conference on Pattern Recognition (submitted)*.

Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.

Peng, Y., Lu, Y.-T., and Chen, Z.-G. (2021). An improved error-based pruning algorithm of decision trees on

large data sets. In *2021 IEEE 6th International Conf. on Big Data Analytics (ICBDA)*, pages 33–37. IEEE.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2:1–10.

Sahatiya, P. (2018). Big data analytics on social media data: a literature review. *International Research Journal of Engineering and Technology*, 5(2):189–192.

Sandhu, A. K. (2021). Big data with cloud computing: Discussions and challenges. *Big Data Mining and Analytics*, 5(1):32–40.

Santana, A., Inoue, S., Murakami, K., Iizaka, T., and Matsui, T. (2020). Clustering-based data reduction approach to speed up svm in classification and regression tasks. In *33rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2020, Kitakyushu, Japan*, pages 478–488. Springer.

Sinanaj, L. (2021). *Allan Variance-based Granulation Technique for Large Temporal Databases*. PhD thesis, University of Massachusetts Lowell.

Sinanaj, L., Haeri, H., Maddipatla, S. P., Gao, L., Pakala, R., Kathiriya, N., Beal, C., Brennan, S., Chen, C., and Jerath, K. (2022). Granulation of large temporal databases: An allan variance approach. *SN Computer Science*, 4(1):7.

ur Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. Y., and Khan, S. U. (2016). Big data reduction methods: a survey. *Data Science and Engineering*, 1:265–284.

Vito, S. (2016). Air Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59K5F.

Yang, Z., Haeri, H., and Jerath, K. (2021). Renormalization group approach to cellular automata-based multi-scale modeling of traffic flow. In Braha, D., de Aguiar, M. A. M., Gershenson, C., Morales, A. J., Kaufman, L., Naumova, E. N., Minai, A. A., and Bar-Yam, Y., editors, *Unifying Themes in Complex Systems X*, pages 17–27, Cham. Springer International Publishing.

Yao, Y., Liu, Y., Yu, Y., Xu, H., Lv, W., Li, Z., and Chen, X. (2013). K-svm: An effective svm algorithm based on k-means clustering. *J. Comput.*, 8(10):2632–2639.

Zhan, Y. and Shen, D. (2005). Design efficient support vector machine for fast classification. *Pattern Recognition*, 38(1):157–161.

Zhang, D.-g., Ni, C.-h., Zhang, J., Zhang, T., Yang, P., Wang, J.-x., and Yan, H.-r. (2022). A novel edge computing architecture based on adaptive stratified sampling. *Computer Communications*, 183:121–135.

Zicari, R. V. (2012). Big data: Challenges and opportunities. *This is Big Data*.