Optimized K-Means Clustering for Enhanced Regression Performance

László Kovács

University of Miskolc,
H-3515 Miskolc-Egyetemváros,
Hungary

Received: 05 July 2025 / Accepted: 30 August 2025 / Published: 25 September 2025

Doi: 10.56345/ijrdv12n2s101

Abstract

In this study, we explore the novel application of clustering techniques in the context of regression analysis. Regression is a key method in data science aimed at predicting the value of real-valued functions for arbitrary input variables. The central hypothesis of our research is that clustering can be leveraged to identify subregions of the input domain where the target function exhibits greater regularity and stability, thus enabling more precise and reliable regression. Specifically, we propose a modified K-means clustering algorithm that optimizes clusters based not only on proximity in feature space but also on the homogeneity of the function values within each cluster, measured via reduced standard deviation. In this paper, we present a series of experimental evaluations analyzing the efficiency of the proposed method. These evaluations also include comparisons with standard k-means clustering and nearest neighbourhood regression methods. The results demonstrate that the proposed method not only reduces the standard deviation within clusters but also improves the regression accuracy.

Keywords: *Regression methods, K-means clustering, Function approximation*

1. Introduction

Clustering is an important data mining tool for data preprocessing. It is used mainly for data reduction, as it generates groups or clusters of objects and the analysis is evaluated on the smaller sized cluster set instead of the large object set. We will focus in this work on the problem domains where the objects are represented by elements in a real vector space. Considering this usual case, there are a few standard clustering methods [1] in the literature, as HAC or K-means to perform the clustering. Beside the data preprocessing task, the clustering methods can be used as stand-alone tools to perform some special data analysis. One important option is outlier detection. In this problem domain, the goal is to detect those objects which are outside the clusters, they stay alone in the object space. The detection of these single objects may have of large practical importance, for example fraud detection, as these objects may represent people with very suspicious behavior.

In our work we focus on the analysis on how we can apply the clustering in solving the regression problems. Regression is also a key data analysis method, where the goal is to predict the function value of real-valued functions for arbitrary input values. Regression is applied in many application fields, like in controlling [2] or forecasting [3]. The simplest case is linear regression where we use linear formulas to predict the target function values. Taking small regions, the linearity assumption can provide a good approximation. Taking more complex approximation schemas, the main problem is the overfitting, i.e. the formulas fitted to a local area cannot extrapolate to larger areas, causing poor prediction accuracy.

The main motivation of our work is the hypothesis that clustering can be used to determine those areas in the input

domain where the target function is more regular, more stable and thus it can provide a more stable approximation. In the next sections, we introduce the key concepts of k-means clustering and then we present our approach on the k-means modification that provides more homogeneous clusters with a reduced standard deviation of the function values. Within these areas, we can achieve better regression precision. After presenting the proposed deviation-driven method, we develop a regression algorithm utilizing this approach. In the test experimental section, we perform different efficiency tests of the different clustering approaches. As the tests results show, the proposed method improves the efficiency of the baseline methods, especially for rare objects spaces, where the input object is far from each other.

2. Related Works

Clustering is typically categorized as an unsupervised learning method, as it operates solely on the input data without requiring labels or additional constraints regarding object similarity. The objects to be clustered are often represented as points in a real-valued vector space $X \subset R^D$, where D denotes the dimensionality of the feature vectors. A significant advantage of using vector space representations is the ability to compute aggregated objects, such as centroids, in a natural and efficient manner. In contrast, when dealing with data from a general metric space, we are limited to a distance matrix without access to explicit vector representations. In such cases, it is generally not possible to compute centroids, which restricts the use of centroid-based clustering algorithms like K-means.

One of the core clustering methods is the Hierarchical Agglomerative Clustering (HAC) [6], where the cluster tree is built up stepwise from bottom up. In the first step, each element is assigned to a separate singleton cluster. The distance between these initial clusters is calculated for each pair of clusters, and the distance values are stored in a distance matrix. In the adjacent iterations, the engine takes the cluster pair with the smallest distance value, and these clusters are merged into a new cluster. Thus, in each iteration step, the number of clusters is reduced by one. Preserving the new clusters and the relationships between the generated clusters, we get a hierarchy of clusters on the input dataset. The constructed cluster hierarchy can be used to visualize the item distribution in the object space, but the method has also some bottlenecks, like

- high computational costs
- inappropriate cluster shape
- unbalanced cluster distribution

Among the various clustering techniques applicable to vector spaces, the K-means algorithm is by far the most widely used due to its simplicity, efficiency, and scalability. K-means was first introduced by Hartigan in 1975 [4], and it partitions a dataset into *K* clusters by iteratively minimizing the within-cluster variance. The big benefit of the K-means method is that it is cost-effective and generates arbitrary cluster shapes.

Having an object set *X* and a cluster count value *K*, the output of the K-means method is a set of centroids $\{c_1, c_2, \dots, c_K\}$ and a partioning of *X* , where

$$\forall i: X_i = \left\{x \in X: \ arg\min_j \{d(x, c_j)\} = i\right\} \quad (1)$$

$$\forall i: \ c_i = \frac{\sum_{x_i \in X_i} x_i}{|X_i|} \quad (2)$$

is met. It can also be shown that the given centroid positions minimize the inter-cluster distances:

$$\sum_i \sum_{x \in X_i} d(x, c_i)^2. \quad (3)$$

The algorithm of K-means can be summarized as follows.

1) Initialization: Select K initial random centroids.
2) Assign each object to the nearest centroid based on Euclidean distance.
3) Recalculate the centroid positions as the mean of all objects assigned to each cluster.
4) Iterate steps 2 and 3 until convergence (no change in assignments of the centroids).

Besides the positive properties, there are some critical points in the K-means method, namely

- It is hard to predict the appropriate cluster count (K)
- The results significantly depend on the selected initial centroid positions.

Regression method [5] is one of the key methods in machine learning, where the main goal is find the best function model providing the most accurate function approximation. Having a real valued function $f_\theta(x)$ with model parameter θ and a training set $T\ \{(x_i, y_i)\}$, the regression procedure determines the optimal function parameter:

$$\theta_o = argmin_\theta \{E(\theta, T)\} \quad (4)$$

where *E* is the error value with í

$$E\ (\theta, T) = \sum_{x \in T} |f_\theta\ (x) - y_x| \quad (5)$$

In most cases, the approximation model is a low-level function of the object features in order to minimize the risk of overfitting. The most widely used approaches are the following methods:

1) Linear regression: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a straight line to the data. It assumes a linear relationship and aims to find the line that minimizes the sum of squared differences between the predicted values and the actual values.
2) Polynomial regression: Polynomial regression models the relationship between variables by fitting a curved line to the data. It's an extension of linear regression that uses a polynomial equation to capture non-linear relationships, allowing the model to fit a wider range of data patterns.
3) Elastic net regression: Elastic Net regression is a regularization technique that combines the penalties of both Lasso and Ridge regression. It's used to improve the stability and performance of linear models by shrinking some coefficients to zero (like Lasso) while also shrinking the remaining non-zero coefficients (like Ridge). This makes it effective for models with many features that may be correlated
4) Regression trees: Regression trees are a type of decision tree used for regression tasks. They split the data into smaller, more manageable groups based on feature values and then predict the output by taking the average of the target variable for all data points within each final group.
5) Regression neural networks: Regression neural networks are a type of artificial neural network designed to predict continuous numerical values. They consist of layers of interconnected nodes that learn complex, non-linear relationships between input features and the target variable, making them powerful for modeling intricate patterns that simpler methods can't capture

All these methods tend to achieve higher accuracy on regular data distributions, where the models more closely align with the true underlying data patterns. The application of clustering on the object set enables the induction of local models on the different clusters instead of a global model on the whole dataset. This local model provides better approximation than the global model. Based on these considerations, our investigation focuses on the development of efficient regular cluster partitions in the object space.

3. Proposed K-Means Method

The K-means clustering method is used to merge the items into groups based only on the position similarities. Thus, the other parameters of the items are irrelevant in the final cluster structure. In the case of regression problems, when the items are assigned to function values, the resulting cluster structure is independent of the function value distribution. Based on this behavior, the clustering cannot help in the improvement of regression efficiency.

To involve the clustering into the regression process, an extended clustering mechanism is required. Our proposal is based on motivation to involve the function values as a new aspect into the clustering process. Thus, the proposed method will involve the function values to measure the similarity or distance between the items.

In the standard K-means method, the optimal position of the centroid is the mean vector of the member objects within the group. Having an object group $X_g \subset X$, the centroid *c* can be given with

$$c = \frac{\sum_{x \in X_g} x}{|X_g|} \tag{6}$$

This formula can be converted into

$$c = c' + \frac{\sum_{x \in X_g} (x - c')}{|X_g|} c = c' + \frac{\sum_{x \in X_g} (x - c')}{|X_g|}, \text{OBJ} \tag{7}$$

where *c'* denotes an arbitrary centroid candidate position.

Hi. In clustering, homogeneity is a metric that evaluates how well a clustering algorithm groups data points that belong to the same true class together. A clustering result is perfectly homogeneous if every cluster contains only data points from a single ground truth class. The Homogeneity Score is a metric that measures the conditional entropy of the class distribution given the cluster assignments. It ranges from 0.0 to 1.0, where a score of 1.0 indicates a perfectly homogeneous clustering result. This score is mathematically defined as:

$$h = 1 - \frac{H(C \mid K)}{H(C)} \tag{8}$$

where:

- H(C) is the entropy of the true class distribution.
- H(C|K) is the conditional entropy of the classes given the cluster assignments. It's low when each cluster is dominated by a single class.

In our investigations, we analyzed two proposed approaches on calculation of the centroid position. The first

method introduces a weighting of the group members objects, while the second method applies an extended feature vector.

In the approach of weighted objects, we give higher relevance for objects (x) having a function value (y_x) more similar with the other's function values. This means that objects with very small or high, outlier values have lower importance in the calculation of the centroid position. Our proposed formula

$$c = \frac{\sum_{x \in X_g} w_x \cdot x}{|X_g|} \quad (9)$$

In order to preserve the validity of Equation (6), we assume that

$$\sum_{x \in X_g} w_i = 1 \quad (10)$$

Regarding the weight values, we applied the following formulas:

$$w'_x = \frac{1}{1+|y_x - \bar{y}|} \quad (11)$$

$$w_x = \frac{w'_x}{\sum {w'}_z} \quad (12)$$

$$\bar{y} = \frac{\sum_{x \in X_g} y_x}{|X_g|} \quad (13)$$

Besides the weighted formula method, we also involved a second approach into the investigations. In this model, we extended the feature vector of the objects with a new component, which stores the related function value:

$$\tilde{x} = < x_1, .., x_D, y_x > \quad (14)$$

This modification maps the related function values to the feature level to be processed at the next level. With this modification of the original representation model, the centroid position will depend not only on the positions of the member objects, but also on the function values of the objects.

One important additional benefit of this method is that it can be easily integrated into the existing baseline K-means implementations. In our approach, the extended feature vector is used only in the centroid construction phase; .in the other modules for evaluation or prediction, we use the original feature vector representation.

4. Efficiency Tests of the Proposed Regression Method

For the implementation of the proposed method, we used the Python framework provided by the Colab engine [7]. For the required calculations we utilized the numpy package.

To evaluate the efficiency and performance of different clustering algorithms, we tested a variety of methods.

1) K1: Baseline K-means method**:** This is the standard K-means algorithm. It randomly initializes cluster centroids and iteratively assigns data points to the nearest centroid, then recalculates the centroids as the mean of the assigned points. This process repeats until the centroids no longer change significantly.
2) K2: Weighted K-means method: An adaptation of the standard K-means algorithm where data points are assigned weights. This allows certain data points, based on the function values, to have a greater influence on the position of their assigned cluster's centroid, which can be useful for handling outliers or emphasizing specific data.
3) K3: Extended K-means method**:** A more advanced variation of K-means that likely incorporates additional features or logic to improve clustering. In this approach, the object position vector is extended with the corresponding function value as a new dimension.

R: Random centroid positioning method: This method serves as a baseline for the impact of centroid initialization. Unlike the iterative K-means methods, this approach simply places cluster centroids randomly without further refinement. Its purpose is to measure the performance of other methods against a purely random, non-iterative approach.

We used two distinct types of synthetic datasets to evaluate the robustness and efficiency of the methods under different data distributions.

- Synthetic uniform data**:** This dataset consists of data points that are randomly and uniformly distributed across a defined space. This type of data presents a challenge for clustering algorithms as there are no inherent clusters, making it a good test for how a method handles noise and identifies arbitrary groupings.

Synthetic blob data: This dataset is designed with distinct, Gaussian-distributed clusters. The data points form clear, spherical "blobs" that are well-separated. This type of data is ideal for testing a clustering algorithm's ability to accurately identify and group the true, underlying clusters.

In the tests, the sum of absolute error was used as the main measure of regression quality. In the experiments, we involved two types of cluster-level approximation formulas:

1) Approximation of the average-value
2) Approximation using a linear regression model

The average version provides a more cost-effective aggregation solution, its is based on a simple aggregation. In the case of linear regression model, the engine calculates the corresponding linear approximation planes. Although this variant seems to be more effective, the experience shows that a more complex formula may cause overfitting, too. Thus, it is important to involve regularization techniques into the model construction procedure.

The presented values in the Figures are aggregated values, for each parameter setting, we executed 5 tests and the calculated average values are shown in the corresponding efficiency figures.

The Figures 1 – Figures 5 present some characteristic dependencies of the proposed methods.

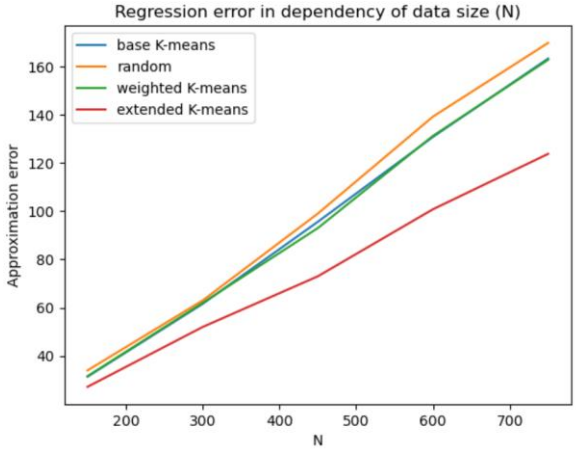

Figure 1. Dependency of the approximation error from the size of the object set (uniform data, avg-approximation, D=5, K=5

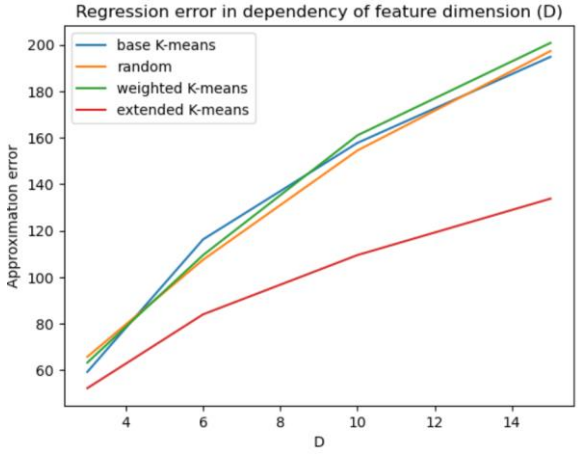

Figure 2. Dependency of the approximation error from the number of dimensions of the object set (uniform data, avg-approximation, N=450, K=5

Figure. 3. Dependency of the approximation error from the number of clusters (uniform data, avg-approximation, N=600, D=5

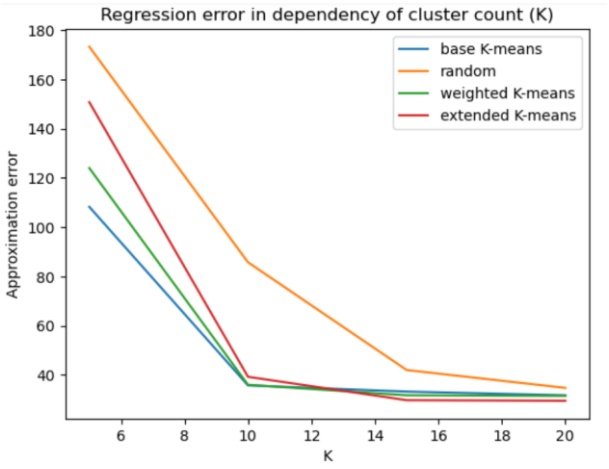

Figure 4. Dependency of the approximation error from the number of clusters (clob data, linear regression-approximation, N=600, D=5

Our evaluation of the test results led to the following important conclusions.

1. The extended K-means approach significantly dominated all the other methods in most of the test's variants.
2. The dominance of the extended K-means method was larger in the tests with uniform data distribution.
3. The weakest solution is related to the random selection method.
4. We could achieve about 25%-30% better prediction accuracy with the proposed extended K-means approach.
5. There was no significant difference in the accuracy comparison between the baseline K-means and the weighted K-means methods.

Figure 5. Dependency of the approximation error from the number of clusters (clob data, average-approximation, N=600, D=5

5. Conclusion

Our research investigated the use of clustering to enhance regression accuracy and model interpretability by employing a cluster-level function approximation approach. We specifically introduced and analyzed two novel methods: the weighted object K-means and the extended K-means. The extensive efficiency tests, which varied parameters like data distribution, number of clusters, and dimensionality, demonstrated the clear superiority of the extended K-means method. This approach consistently outperformed all other methods, achieving a remarkable 25% to 30% improvement in prediction accuracy. This finding strongly validates our hypothesis: segmenting data with specialized clustering techniques and applying local regression models is a highly effective strategy. The extended K-means method not only provides a powerful means to boost predictive performance but also supports the development of more granular and explainable models.

References

Colab framework, https://colab.research.google.com/ (2025)

Gutiérrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., & Hervas-Martinez, C. (2015). Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 127-146.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100-108..

Higgins, J. P., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in medicine*, *23*(11), 1663-1682.

Kaur, P. J. (2015, March). A survey of clustering techniques and algorithms. In *2015 2nd international conference on computing for sustainable global development (INDIACom)* (pp. 304-307). IEEE.

Monath, N., Dubey, K. A., Guruganesh, G., Zaheer, M., Ahmed, A., McCallum, A., ... & Wu, Y. (2021, August). Scalable hierarchical agglomerative clustering. In *Proceedings of the 27th ACM SIGKDD Conference on knowledge discovery & data mining* (pp. 1245-1255).

Zubakin, V. A., Kosorukov, O. A., & Moiseev, N. A. (2015). Improvement of regression forecasting models. *Modern applied science*, *9*(6), 344-353.