Scalable Clustering of Amazon Book Reviews using the BFR Algorithm

Alex Tulip

Department of Computer Science, University of Milan, Italy.

Abstract

This report presents a scalable clustering system implemented from scratch to analyze the Amazon Books Review dataset. To handle the massive size of the dataset (3GB+), we utilized the Bradley, Fayyad, and Reina (BFR) algorithm, a variant of K-Means designed for disk-resident data. The system processes data in chunks, maintaining statistical summaries ($\boldsymbol{N, SUM, SUMSQ}$) to perform clustering in a single pass over the data. Our analysis identifies $\boldsymbol{K = 4}$ distinct market segments in the book industry, separating products not just by price, but by a non-linear relationship between popularity and rating.

Keywords: Clustering, BFR Algorithm, Massive Datasets, Data Mining, Python

1 Introduction and Dataset

The objective of this project is to uncover latent structures in the book market using unsupervised learning techniques capable of scaling to massive datasets.

The analysis is based on the **Amazon Books Review** dataset hosted on Kaggle. The raw dataset consists of over 3 million reviews (approx. 3GB). Due to memory constraints, loading the entire dataset into RAM is infeasible. Therefore, we access the data using a streaming approach, processing the CSV file in chunks of 100,000 rows.

1.1 Data Organization

The raw data contains individual reviews. To perform meaningful clustering, we aggregated these reviews to create a profile for each unique book. We utilized a dictionary-based hash map to stream the data and update the following sufficient statistics for each book title:

- **Count:** Total number of reviews.

- **Sum of Ratings:** To compute the average rating.
- **Price Information:** To compute the average price (handling metadata inconsistencies).

This aggregation reduced the dataset from millions of review rows to approximately 48,000 unique book entities, which serve as the data points for our clustering algorithm.

2 Preprocessing

The BFR algorithm operates in a Euclidean space and relies on the Mahalanobis distance. This requires features to be numerical and comparable. We selected three features for clustering:

1. **Price:** The cost of the book.
2. **Average Rating:** A value between 1 and 5.
3. **Review Count:** A proxy for popularity.

2.1 Normalization

These features operate on vastly different scales (e.g., Price ≈ 20, Rating ≈ 4, Count $\approx 10^3$). To prevent the variance of the "Review Count" feature from dominating the distance calculations, we applied **Z-score normalization** to all features:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

This ensures that all dimensions contribute equally to the cluster definition, a critical requirement for the Mahalanobis distance used in BFR.

3 Algorithms and Implementation

We implemented the **BFR (Bradley, Fayyad, and Reina) Algorithm** entirely from scratch in Python. BFR is a variation of K-Means specifically designed for high-dimensional data that cannot fit in main memory.

3.1 BFR Architecture

Our implementation manages three distinct sets of points, as defined in the course literature:

- **Discard Set (DS):** Points that are close enough to a cluster centroid to be summarized. These are represented only by their statistics ($N, \mathbf{SUM}, \mathbf{SUMSQ}$) and the points themselves are discarded.
- **Compressed Set (CS):** Points that are outliers relative to the main clusters but are close to each other. These form "miniclusters" and are also summarized.
- **Retained Set (RS):** True outliers that fit neither into the DS nor the CS. These are kept in memory as individual points.

3.2 Mahalanobis Distance and Variance Fix

Points are assigned to the Discard Set if their Mahalanobis distance to a centroid is below a threshold (set to 3.0).

$$d(\mathbf{x}, \mathbf{c}) = \sqrt{\sum_{i=1}^{d} \left(\frac{x_i - c_i}{\sigma_i} \right)^2} \tag{2}$$

Implementation Challenge: During initialization, clusters formed with a single point have a variance of zero. This causes the Mahalanobis distance to become undefined (division by zero), preventing the cluster from growing. **Solution:** We implemented a heuristic where clusters with $N < 10$ utilize a default standard deviation of $\sigma = 1.0$ (matching the normalized global variance). This allows new clusters to "warm up" and absorb their initial neighbors before switching to their calculated internal variance.

4 Scalability

The proposed solution scales linearly $O(N)$ with the size of the dataset. By processing the raw CSV in chunks and strictly maintaining summary statistics, the memory footprint of the algorithm is determined by the number of clusters K, not the number of data points N. The aggregation step reduces the data volume significantly, and the BFR logic ensures that even if the number of books were to increase to millions, the memory usage would remain stable as points are continuously discarded into summary vectors.

5 Experiments

We determined the optimal number of clusters K using the Elbow Method, testing values $K \in [2, 6]$. We ran the full BFR pipeline for each K and calculated the Sum of Squared Errors (SSE) for the Discard Set.

As shown in Fig. 1, the SSE decreases as expected from $K = 2$ to $K = 4$. However, at $K = 5$, the error increases significantly. This "inverted elbow" indicates that forcing more than 4 clusters destabilizes the model, causing the BFR algorithm to expand cluster variances excessively to absorb points, or rejecting too many points into the Retained Set. Thus, $K = 4$ was selected as the optimal configuration.

6 Results and Discussion

The clustering results for $K = 4$ reveal distinct market segments in the book industry. Fig. 2 visualizes these segments.

The four identified clusters are interpreted as follows:

1. **Cluster 0 (Expensive/Niche):** High price ($\approx$ \$31), perfect rating (5.0), but very low review counts. These represent niche text-books or collectors' items.

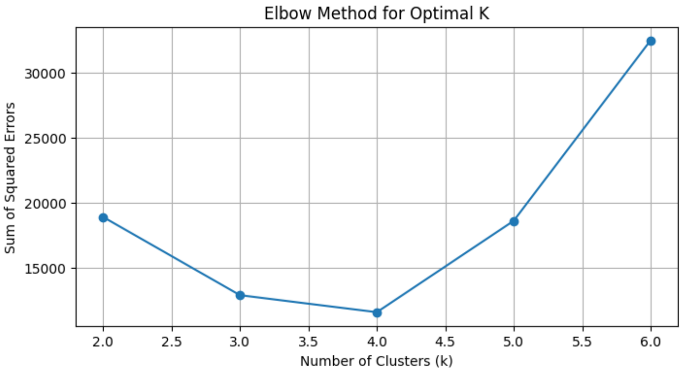

Fig. 1 Elbow Method showing the SSE for different K values. Note the distinct minimum at K=4 and the instability (SSE increase) at K=5 and K=6.

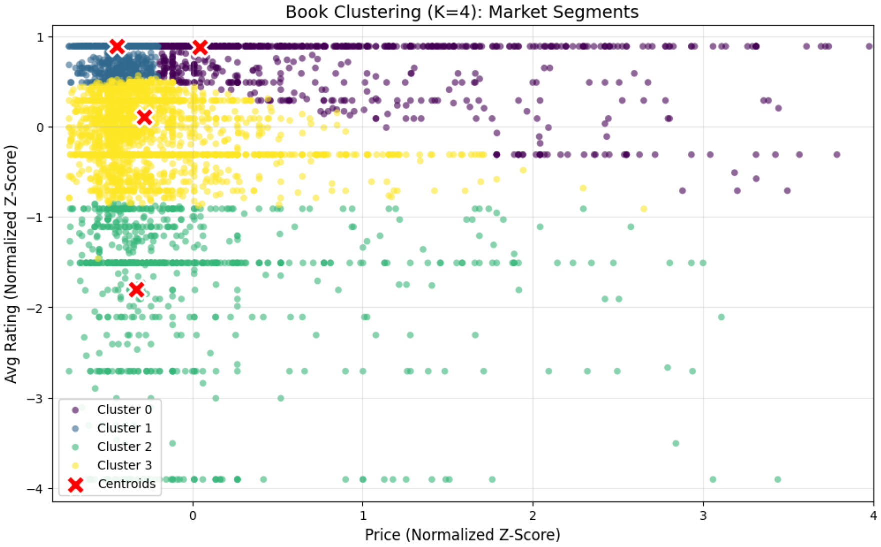

Fig. 2 Visualizing the 4 Market Segments (K=4). The zoomed view focuses on the main density of data, with the centroids marked by red crosses.

2. **Cluster 1 (Cheap/Top-Rated):** The largest cluster. Low price (≈ $12) with 5-star ratings. These are the ”hidden gems” of the catalog.
3. **Cluster 2 (The Flops):** A distinct group defined by low ratings (≈ 2.7 stars). BFR successfully separated these underperforming products from the generally high-rated baseline.
4. **Cluster 3 (Mainstream):** Characterized by the highest review counts (avg 7) and solid ratings (4.3 stars). These are popular, widely-read books.

The vertical alignment of centroids in Fig. 2 highlights that for the majority of books (Clusters 1, 2, 3), price is not a differentiating factor; the market segments are defined primarily by quality (Rating) and popularity (Review Count).

Declaration of Authorship

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study. No generative AI tool has been used to write the code or the report content.