# 1 Performance of PKBD Clustering on Real World Datasets

## Introduction

We have applied the PKBD clustering to the following datasets that reflect variety in terms of sample size, dimensions, and number of clusters. In what follows, we provide a description of the various datasets, describe in detail the preprocessing steps, and provide the results. The associated datasets and scripts can be accessed in the GitHub repository at `https://github.com/rmj3197/QuadratiK-Performance`.

1. Birch Dataset II [6]: This dataset contains $300,000 \times 2$ dimensional synthetic data vectors that represent one of the three patterns shown in Figure 1. The dataset can be found at - `https://cs.joensuu.fi/sipu/datasets/`.

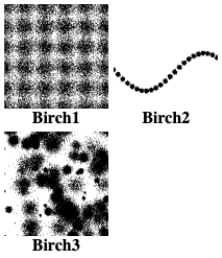


Figure 1: The vectors in the Birch dataset represent one of the three clusters. Image is taken from `https://cs.joensuu.fi/sipu/datasets/`.

2. 20 Newsgroups Text Dataset: This dataset is a part of the `scikit-learn` package [5]. This dataset is a collection of approximately 18,000 news posts, partitioned (nearly) evenly across 20 different newsgroups. More information of the dataset can be found at: `http://qwone.com/~jason/20Newsgroups/`.

3. Reddit Text Dataset: This text dataset is a part of Massive Text Embedding Benchmark (MTEB) [3]. This dataset contains titles of Reddit posts from subreddits. For this work, we use the subset of the "test" set of the data (data split number 13 out of 25) available from `https://huggingface.co/datasets/mteb/reddit-clustering`.

4. StackExchange Text Dataset: This dataset is also part of the MTEB. This dataset contains titles of questions posted on various StackExchanges. The clustering is performed to identify the different StackExchange communities from which the question titles are collected. In this work, we have used a subset of the "validation" set (data split number 9 out of 25) of

the dataset. The dataset can be found at `https://huggingface.co/datasets/mteb/stackexchange-clustering`.

5. 3 Newsgroups Text Dataset: In this dataset we specifically select three groups from the 20 Newsgroups Text Dataset. The groups selected are: `omp.os.ms-windows.misc`, `rec.sport.hockey`, and `soc.religion.christian`, which represent news related to the operating system MS Windows, Hockey, and Christianity. We apply the clustering algorithm just to a subset of the data consisting of these three groups.

6. arXiV Text Dataset: This dataset was introduced in [1] and is now also part of the MTEB. This dataset was originally curated for classification purposes, where 10,000 words from papers were extracted to train classifiers to predict the paper category among 11 arXiv classes. In our work, we use the "test" set of the data obtained from `https://huggingface.co/datasets/mteb/ArxivClassification`.

7. Hand-written Digits Image Dataset: This $8 \times 8$ pixels image dataset is introduced in [2]. This dataset contains images of handwritten digits from 0-9, representing the 10 classes where each class refers to a digit. This dataset is available in the `scikit-learn` package and also available on UCI ML Repository at `https://archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits`. Representative images from the various classes are shown in Figure 2.

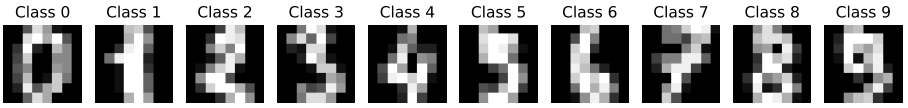


Figure 2: Example images for each digit from the Hand-written Digits Image Dataset.

8. Detect AI Generated vs Student Generated Text Dataset: This dataset is taken from Kaggle at `https://www.kaggle.com/datasets/prajwaldongre/llm-detect-ai-generated-vs-student-generated-text`. The dataset consists of text written by either a student or a large language model (LLM). The goal is to identify the clusters representing the LLM-generated and the student-written text.

### Data Preprocessing

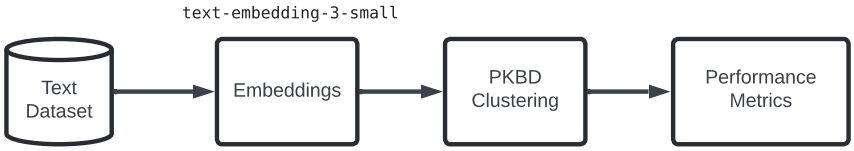


Figure 3: Pipeline used for clustering text datasets.

For clustering text datasets, we have used the workflow depicted in Figure 3. The text is first converted into embeddings using the `text-embedding-3-small` embedding model provided by OpenAI. More information on the text embedding model can be found in [4]. The obtained embeddings are used as the data vector to perform the PKBD clustering. The `text-embedding-3-small` embedding model used in our work supports a maximum of 8191 tokens. We removed any text containing five or fewer tokens. Additionally, we set a maximum limit of 7000 tokens. If a text exceeded this limit, it was trimmed to fit within the constraint; otherwise, it was kept as is. The tokenization was performed using the `cl100k_base` tokenizer in `tiktoken Python` library. Embeddings of different sizes are used for various datasets shown to demonstrate that the PKBD clustering algorithm can also be applied to high-dimensional data. In the case of the image dataset, the pixel values are flattened, used for clustering.

## 2 References

# References

[1] Jun He, Liqun Wang, Liu Liu, Jiao Feng, and Hao Wu. Long document classification from local word glimpses via recurrent attention learning. *IEEE Access*, 7:40707–40718, 2019.

[2] Cenk Kaynak. Methods of combining multiple classifiers and their application to handwritten digit recognition. Master's thesis, Fen Bilimleri Enstitüsü, 1995.

[3] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[4] OpenAI. Vector Embeddings - OpenAI API, 2025. [Online; accessed 2025-03-01].

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.