

A Survey on Customer Segmentation and Natural Language Tasks

RBCDSALIITM

N Kausik ¹ Dr. B. Ravindran ¹
¹IIT Madras

Introduction

As companies and businesses grow and expand, the data they obtain and generate increases. They use it to make data-driven business decisions that also require the use of various algorithms, machine learning and AI techniques. Over the years there have been numerous algorithms and research published on all the problems, however for any uninformed person it is very difficult to navigate through all the different options to solve their business problem effectively. Through this project, we address this issue by developing a set of rules and decisions to select the best algorithm or model for two areas, namely Customer Segmentation and Natural Language Tasks (8 tasks). The rules are condensed into easy-to-use flowcharts so that anyone with a business need can easily obtain the references, algorithms or models best suited for their needs.

Methodology

For the tasks, the following process was followed to obtain the flowcharts and references,

- First, we documented available state-of-art research papers, datasets, evaluation metrics, articles, GitHub repositories and other resources for the task.
- Then, we implemented and evaluated the various models or algorithms on popular datasets and documented the results.
- Observations were made based on the evaluations and properties of the algorithms or models and a flowchart was constructed based on them. Applications were also built.

Customer Segmentation

- Clustering algorithms can be broadly classified into Density-Based, Learning-Based, Hierarchical or Graph-Based algorithms. Further, some algorithms perform soft clustering which provides the membership distribution for a point over the clusters instead of assigning just one cluster to it.
- Evaluation metrics like Silhouette Score, Calinski-Harabasz Index and Davies-Bouldin Index were used to compare the algorithms.
- The flowchart was constructed using the properties, observations and metrics of the algorithms on 5 popular customer segmentation datasets.
- An application was built to run the clustering algorithms on datasets.

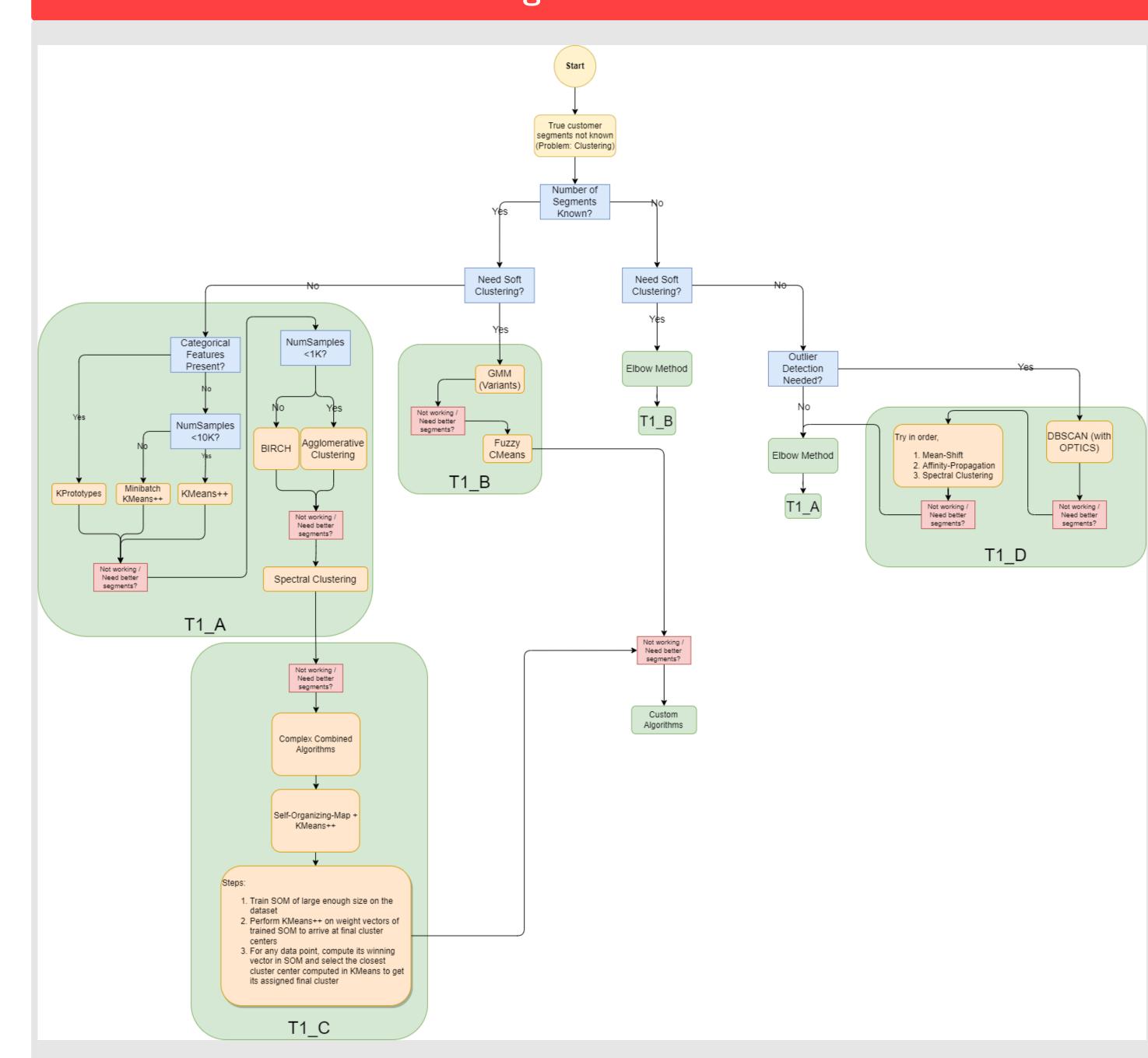
Natural Language Tasks

- The 8 tasks surveyed are Sentiment Analysis, Named Entity Recognition, POS Tagging, Dialogue, Summarisation, Translation and Question Answering.
- For each task, we documented popular datasets and state-of-art models from PapersWithCode and NLPProgress.
- Further, available models from the Hugging-Face platform were also collected and various aspects like their related dataset, metrics, etc were documented.
- The collected Hugging-Face models were evaluated on available Hugging-Face datasets using their library on Google Colab Notebooks (CPU only). The obtained results and metrics (performance and time) were documented.
- The flowchart for each task was constructed using the properties, observations and metrics of the models.
- Google Colab notebooks were made for various use cases like using the model, evaluating and fine-tuning for the various tasks.
- An application was made for using available Python libraries to solve the tasks.
- Another application was made for tabulating and comparing the Hugging-Face and state-of-art models using the documented information.

Results

All the obtained results, documentation and code can be found at https://github.com/KausikN/MTech_MTP_TextProblems. All the flowcharts for language tasks follow the same structure and hence only the flowchart of sentiment analysis is shown.

Customer Segmentation Flowchart



Sentiment Analysis Model Selection

