Spectral Team Report

December 1, 2023

Abstract

Our team has collected four datasets on academic publications from the math faculty, including citations, journals, MSC codes, and references. Each dataset represents professors with their publication statistics. Our preprocessing involved removing columns with only a single non-zero entry, as they uniquely correspond to individual professors and don't contribute to clustering. We applied L1 and L2 normalization to mitigate data skewness, followed by PCA for dimensionality reduction. Our exploration of clustering methods, including K-means, Affinity Propagation, and Spectral Clustering, revealed that K-means on the MSC dataset provided the most coherent clusters. This suggests that MSC codes are effective in categorizing faculty research areas.

1 Dataset Introduction

In this project, we analyze four distinct datasets, each encompassing data related to 69 professors from the Mathematics Department. These datasets – Journals, References, Citations, and MSC – offer a comprehensive view of the professors' academic contributions and interactions within the field.

Journals Dataset: This dataset catalogs the publication records of the professors across various academic journals. Each column is named after a specific journal, such as "Annals of Combinatorics" or "Mathematische Annalen". The entries under each column signify the number of times a professor (represented by each row) has published in that particular journal.

MSC Dataset: The MSC (Mathematics Subject Classification) dataset categorizes the professors' publications according to MSC codes. Each column represents a different MSC code. The entries denote the count of papers that a given professor (each row) has published under each MSC code.

Citations Dataset: This dataset focuses on the impact of the professors' work through citations. Columns in this dataset represent individual papers, identified by unique identifiers. The entries indicate the number of times a paper (in a specific column) has cited any work by the professor corresponding to a particular row.

References Dataset: Contrasting with the Citations dataset, the References dataset tracks the professors' engagement with existing literature. Like the Citations dataset, its columns represent papers by unique identifiers. However, the entries here reflect the number of times a professor (each row) has cited a specific paper (each column).

Together, these datasets provide a multifaceted view of the academic output and influence of the Mathematics Department's faculty, capturing both their contributions to and interactions with the broader mathematical community.

2 Data Preprocessing

Given the characteristics of our datasets, particularly in the Citations and References datasets, we observed a common occurrence: numerous columns, representing papers by their unique identifiers, contained only a single non-zero entry. This implies that these papers have limited interaction or influence in relation to the majority of the professors in our study. Such instances of minimal engagement do not contribute significantly to the analysis of academic interactions within the department.

To address this, we implemented a masking technique. This method involves excluding columns that have only one non-zero entry across all datasets. The rationale behind this approach is to focus our analysis on more impactful data, where the interactions or influences are more pronounced and widespread.

Then, we employed two different normalization techniques, L1 and L2 normalization, to create two distinct versions of each dataset.

L1 Normalization: Also known as least absolute deviations, this technique adjusts the values in a dataset so that the sum of the absolute values in each row equals one. It does this by dividing each value by the sum of the absolute values in its row. This approach is beneficial in our project as it helps in reducing the impact of outliers and ensuring that extreme values in the data do not disproportionately affect the overall analysis. L1 normalization enhances the robustness of our models against anomalies and offers a clear interpretation in terms of relative proportions of each professor's contribution or interaction.

L2 Normalization: Often referred to as least squares, L2 normalization works by dividing each value in a row by the square root of the sum of the squared values in that row. The result is that the sum of the squares of the values in each row is one. L2 normalization is particularly useful in our context as it preserves the geometric relationships in the data, such as the Euclidean distances between points (or professors, in our case). This is crucial for applications where these relationships are meaningful, such as in clustering or principal component analysis (PCA).

By applying both L1 and L2 normalization to our datasets, we aim to leverage their distinct advantages. L1 normalization provides a robust model less sensitive to outliers, which is particularly beneficial when dealing with skewed data or anomalies. On the other hand, L2 normalization preserves data relationships, making it ideal for analyses involving distances or correlations. The use of both normalization techniques allows us to compare and contrast the outcomes and insights derived from each method, thereby enriching our understanding of the academic interactions within the Mathematics Department.

In the subsequent phase of our analysis, we employed Principal Component Analysis (PCA) to address the challenge of high dimensionality in our Citations and References datasets, each comprising over 30,000 columns. PCA is a sophisticated statistical technique designed to transform a complex dataset into a simpler structure without significantly losing the essence of the original data.

Algorithm of PCA: PCA begins by standardizing the data, ensuring each feature contributes equally to the analysis. It then computes a covariance matrix to understand the relationships between different features. Following this, it involves the calculation of eigenvectors and eigenvalues from this covariance matrix. Eigenvectors, representing the directions of the new feature space, are paired with eigenvalues, which indicate the magnitude of variance along these directions. The dataset is then projected onto a smaller space defined by the most significant eigenvectors, known as principal components. These components are selected based on the magnitude of their corresponding eigenvalues, with higher values indicating more significant components.

Advantages in Dimensionality Reduction: The application of PCA offers several benefits. Firstly, it reduces the risk of overfitting by decreasing the number of features. Secondly, it facilitates data visualization and interpretation by reducing dimensions to a more manageable level. Thirdly, PCA retains the most significant variance features of the dataset, ensuring that the most vital information is preserved. Furthermore, it enhances computational efficiency and can assist in noise reduction, focusing analysis on the most informative aspects of the data.

By applying PCA to our datasets, we effectively reduced their dimensionality. This step was crucial in simplifying the datasets for more focused and efficient analysis, especially given the high number of features originally present in the Citations and References datasets.

3 Methods

After cleaning and preprocessing the datasets, we employed two distinct clustering techniques from the scikit-learn library: Affinity Propagation and K-Means.

Affinity Propagation: Affinity Propagation is a clustering algorithm that does not require the number of clusters to be specified in advance. It operates by sending messages between pairs of samples until a set of exemplars emerges. These exemplars are representative data points that best describe the data. The algorithm has two main steps:

Responsibility: The responsibility message reflects the suitability of candidate exemplar point k for data point i, considering other potential exemplars for point i.

Availability: The availability message reflects the appropriateness of point i choosing point k as its exemplar, considering the support from other points that point k should be an exemplar. The messages are updated iteratively, and the process converges to a set of exemplars and corresponding clusters. This method is particularly useful for complex datasets since it does not require pre-specification of the number of clusters.

K-Means: K-Means is a popular clustering method that partitions the dataset into K clusters, where each data point belongs to the cluster with the nearest mean. The algorithm follows these steps:

Initialization: Select K random points as cluster centers (centroids).

Assignment: Assign each data point to the nearest centroid, forming K clusters.

Update: Recalculate the centroids as the mean of all points assigned to the cluster.

Iterate: Repeat the assignment and update steps until convergence (when assignments no longer change).

To determine the optimal number of clusters K in K-Means, we used the Elbow Method. This involves running the K-Means algorithm on the dataset for a range of values of K (e.g., from 1 to 30) and calculating the sum of squared distances from each point to its assigned center. When these overall intra-cluster distances are plotted against the number of clusters, the 'elbow' point where the rate of decrease sharply changes represents an appropriate number of clusters. The idea is that adding another cluster does not give much better modeling of the data after this point.

By applying these clustering techniques to our datasets, we aim to uncover meaningful patterns and groupings inherent in the data, facilitating a deeper understanding of the underlying structures within our academic datasets.

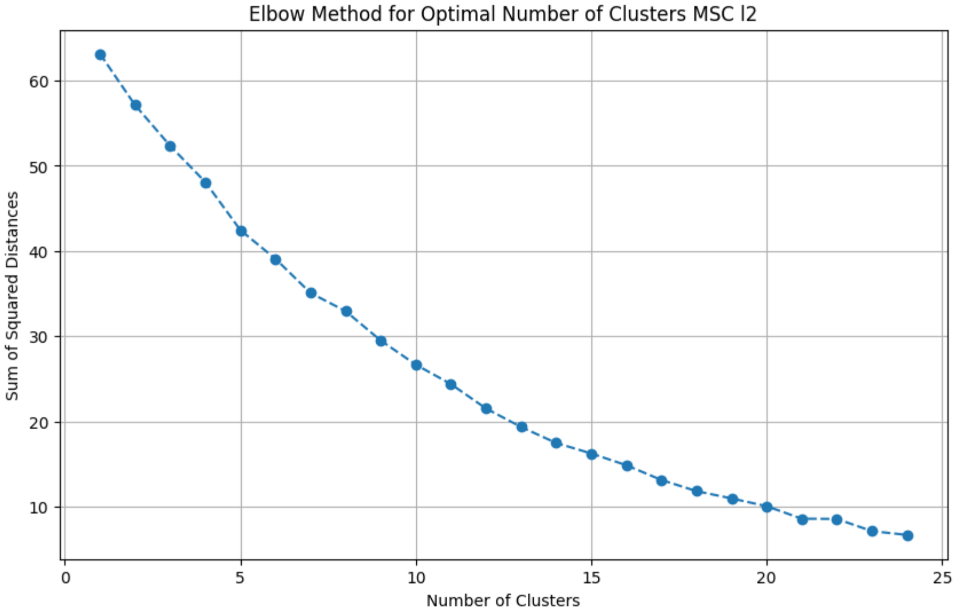

Figure 1: Elbow Plot Analysis

4 Results and Future Developments

Since most of our clusterings are all over 10 subgroups, it's hard to visualize all the results here. Below is the result clustering with l2 normalized MSC dataset with 30 PCA components number, and 15 K-means clustering number.

- **Cluster 0**: DeVilleRELee, RaptiZoi.
- **Cluster 1**: BradlowStevenBenjamin, HaboushWilliamJ, HellerJeremiahBen, JandaFelix, KatzSheldonH, YongAlexanderTF.
- **Cluster 2**: AlbinPierre, FernandesRuiLoja, HungPei-Ken, KermanEly, LermanEugeneM, NikolaevIgorG, PascaleffJamesThomas, TolmanSusan.
- **Cluster 3**: AhlgrenScottD, BocaFlorin-Petre, DuursmaIwanMaynard, FordKevinB, ReznickBruce, ThornerJesse, ZaharescuAlexandru.
- **Cluster 4**: DiFrancescoPhilippe, KirkpatrickKayLene, LeditzkyFelix, YoungAmanda.

- **Cluster 5**: DunfieldNathanM, GuzmanRosemaryK, RasmussenJacob, Rasmussen-SarahDean.
- **Cluster 6**: BaryshnikovYuliyM, DeyParthaSarathi, SongRenming, SowersRichardB, WeiWei, WuXuan.
- **Cluster 7**: CooneyDanielB, FadinaTolulope, FengRunhuan, JingXiaochen, QuanZhiyu.
- **Cluster 8**: BronskiJaredC, ErdoğanMehmetBurak, HurVeraMikyoung, KirrEduard, LaugesenRichardSnyder, LiXiaochun, TzirakisNikolaos, ZharnitskyVadim.
- **Cluster 9**: BaloghJózsef, KostochkaAlexandrV.
- **Cluster 10**: JungeMarius, KutzarovaDenkaN, OikhbergTimur.
- **Cluster 11**: IvanovSergeiVladimirovich, LiuYuan, MineyevIgor.
- **Cluster 12**: Berwick-EvansDaniel, HiraniAnilN, McCarthyRandy, RezkCharlesW, StojanoskaVesna.
- **Cluster 13**: HinkkanenAimo, TysonJeremyT.
- **Cluster 14**: DoddChristopher, DuttaSankarPrasad, KedemRinat, LaNaveGabriele, TumanovAlexander.

But we have uploaded all our clustering results as CSV files for further analysis. During our analysis process, the MSC dataset emerged as the most effective in categorizing faculty, while other datasets like journals or citations showed potential biases. So far, our limitation has been the inability to find an effective way to combine all the different academic publication statistics together. We are only able to analyze them separately. A future direction could involve improving the preprocessing of these datasets and exploring methods to integrate them effectively with varying weights.