Knowledge Discovery in Databases: An Exploration of YouTuber Data

Sanjay Bhargav Kudupudi

September 28,2023

Abstract

This paper elucidates the Knowledge Discovery in Databases (KDD) process through a structured exploration of a dataset encompassing various attributes of YouTubers. We demonstrate the applicability and effectiveness of each stage of the KDD process, emphasizing the importance of preprocessing and the potential of clustering algorithms in uncovering patterns in digital media datasets.

1 Introduction

In the era of Big Data, extracting meaningful insights from vast datasets has become pivotal. The Knowledge Discovery in Databases (KDD) process provides a systematic approach to this endeavor. This research employs a dataset on YouTubers to exemplify each step of the KDD process.

2 Methodology

The dataset comprises multiple attributes, including username, content categories, subscriber count, country of origin, and engagement metrics. Our exploration followed the standard KDD process, encompassing the following steps:

- Understanding the Domain
- Data Selection
- Data Preprocessing

- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Representation

3 Results and Discussion

3.1 Data Preprocessing

Missing data in the 'Categories' column was addressed by imputation, assigning the placeholder "Unknown". Outliers, especially among popular YouTubers, were identified and retained due to their validity.

3.2 Data Transformation

Log transformation was applied to numeric columns, enhancing the normality of distributions, thus making the dataset more amenable to various analytical techniques.

3.3 Data Mining

The KMeans clustering algorithm, guided by the elbow method, revealed that three clusters provided an optimal balance between precision and computational cost. This classification potentially segments YouTubers into distinct tiers based on metrics such as popularity and engagement.

4 Conclusion

The KDD process offers a robust approach to data exploration and knowledge extraction. Through our investigation of YouTuber data, we underscored the significance of rigorous preprocessing and the promise of clustering in discerning patterns within digital media datasets.

5 Acknowledgments

We express our gratitude to all contributors and experts who provided the dataset and tools essential for this research.

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, 17(3):37, 1996.

[2] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 3rd edition, 2011.

[3] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 4th edition, 2016.

[4] Soumen Chakrabarti, Byron Dom, and Piotr Indyk. *Enhanced hypertext categorization using hyperlinks*. In ACM SIGMOD Record, volume 27, pages 307–318. ACM, 1998.

[5] Ronald J. Brachman and Tomasz Anand. *The Process of Knowledge Discovery in Databases*. In Advances in Knowledge Discovery and Data Mining, pages 37–57. AAAI/MIT Press, 1996.