Enhancing E-Commerce Store Clustering Using Frequent Itemset Mining and Mixed Data Analysis

Dat Nguyen[1], Huong Bui[1]

[1]Faculty of Information Technology, HUTECH, Ho Chi Minh City, Vietnam

April 3, 2025

Overview

1. INTRODUCTION
2. PROPOSED METHOD
3. EXPERIMENTAL RESULTS
4. CONCLUSION

1. INTRODUCTION

- Our research clarified 2 main issues:
  - Clustering stores on the Tiki e-commerce platform to support the platform in optimizing business performance and managing more effectively.
  - Maximal Frequent Itemset Mining (FP-Max) is applied to enrich data features, then integrated with clustering methods to improve the clustering performance.
- **Keywords:** *E-commerce Store Clustering, Maximal Frequent Itemset Mining, Mixed Data Analysis.*

2. PROPOSED METHOD

Data Collection:

- Data collected on the Tiki.vn e-commerce platform, including 933 stores & 9 features.

Data Preprocessing:

- **Data Cleaning:** Remove invalid/missing values.
- **Feature Selection:** Wilson score interval for feature combination. Eliminate highly correlated features.
- **Normalization:** StandardScaler.
- **Noise Handling:** Winsorization, Isolation Forest, and DBSCAN.

Revenue	YearJoined	Followers	ChatResponse	RatingQuality	PositiveQuality
509,781,800	6	1717	1	4.536446274	0.7506283509
205,821,900	5	476	0.83	4.764938241	0.8300109782
130,629,912	4	8504	0.7	4.458951116	0.7702080451

Table: Data Description

Research Workflow:

- **Step 1:** Perform initial clustering to establish evaluation baseline.
- **Step 2:** Create new binary features using FP-Max.
- **Step 3:** Integrate these features & initial data to improve clustering results.

2. PROPOSED METHOD

Step 1: Perform initial clustering to establish evaluation baseline.

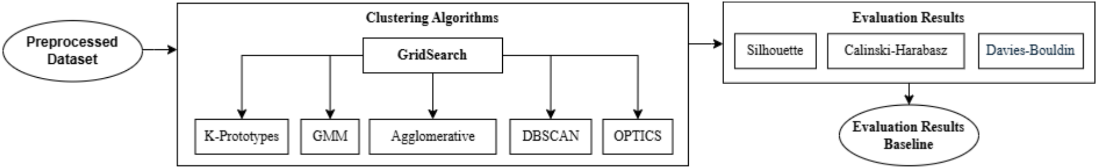

Figure: Initial clustering process

Models	Number of Clusters	Parameters	Silhouette score
K-Means	5	init = *"k-means++"*	0.3254
Agglomerative	4	linkage = *"ward"*	0.3078
GMM	3	covariance_type = *"tied"*	0.3284
DBSCAN	2 (noise = 0)	eps = *"1.1"*, minpts = *"12"*	0.3351
OPTICS	6 (noise = 709)	min_sample = *"12"*, xi = *"0.05"*, min_cluster_size = *"10"*	0.5142

Table: Hyperparameter Tuning Results

Step 2: Create new binary features using FP-Max.

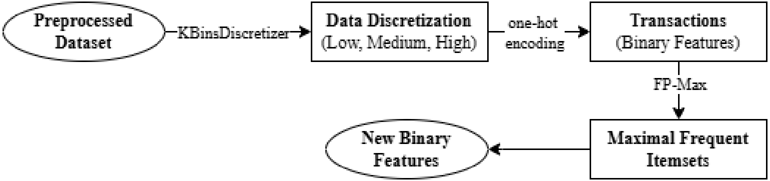

Figure: Create new binary features process

RatHigh_revLow_PosHigh	FolLow_revLow	ChatHigh	PosHigh_revLow_ChatLow	...
1	1	0	1	...
0	0	0	0	...
1	0	1	0	...
...	...	...	...	...

Table: New binary features created

Step 3: Integrate these features & initial data to improve clustering results.

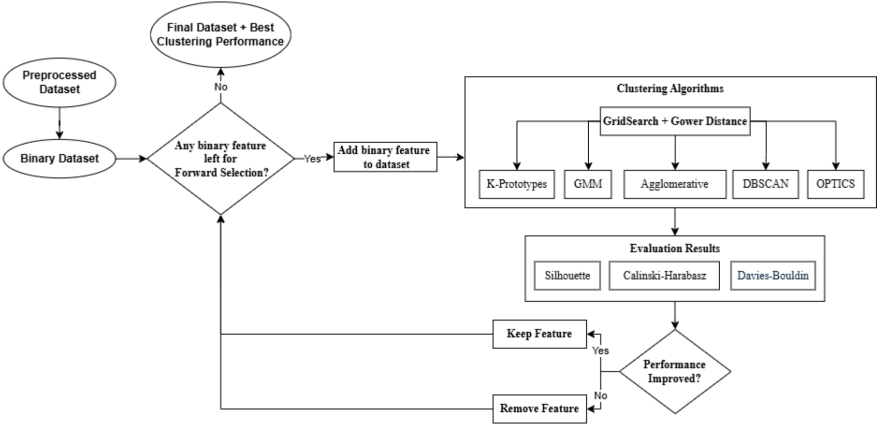

Figure: Integrate new binary features process

3. EXPERIMENTAL RESULTS

Comparison of Clustering Results Before and After Integrating Binary Features:

- **Binary Features Added:**
  - FollowersLow_YearJoinedLow_RevenueLow_ChatLow.
  - YearJoinedLow_PositiveHigh.
- **Evaluation Metrics:**
  - Silhouette (Higher is better)
  - Davies-Bouldin (Lower is better)
  - Calinski-Harabasz (Higher is better)
- **Effective Clustering Methods:**
  - Works well for Agglomerative, K-Means/K-Prototypes, and GMM.
  - Does not work well for DBSCAN and OPTICS.

Algorithms	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.325	1.088	415.934
Agglomerative	0.308	1.036	361.240
GMM	0.328	1.193	362.205
DBSCAN	0.335	1.274	427.131
OPTICS	0.514	0.662	270.900

Table: Initial clustering evaluation results.

Algorithms	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Prototypes	0.459 ↑	0.963 ↓	495.393 ↑
Agglomerative	0.516 ↑	0.959 ↓	994.749 ↑
GMM	0.340 ↑	1.242 ↓	367.991 ↓
DBSCAN	- ↓	- ↑	- ↓
OPTICS	- ↓	- ↑	- ↓

Table: Clustering results after integrating binary features.

Cluster Analysis:

- K-Prototypes was selected for its high cluster number and strong performance.
- High statistics & low p-values indicate significant cluster differences.

Feature	Test	Statistic	p-value
Revenue	ANOVA	36.88	1.92e-28
YearJoined	ANOVA	522.09	2.22e-226
Followers	ANOVA	30.59	8.46e-24
ChatResponse	ANOVA	3221.11	0.00
RatingQuality	ANOVA	140.20	7.51e-92
PositiveQuality	ANOVA	255.01	1.60e-143
FollowersLow, YearJoinedLow, RevenueLow, ChatLow	Chi-squared	543.55	2.54e-116
YearJoinedLow, PositiveHigh	Chi-squared	330.26	3.21e-70

Table: Statistical Test Results Summary

Cluster Analysis:

- **Cluster 0**: Long-standing, low-efficiency stores → Needs improvement in product, service, and engagement.
- **Cluster 1**: Newer stores with good but inconsistent performance → Optimize strategies and promotions.
- **Cluster 2**: Inefficient long-standing stores with some positive scores → Focus on operations product innovation.
- **Cluster 3**: New stores with average performance → Growth via training, sales support, and better service.
- **Cluster 4**: Highly efficient, long-standing stores → Retain and support with marketing partnerships.

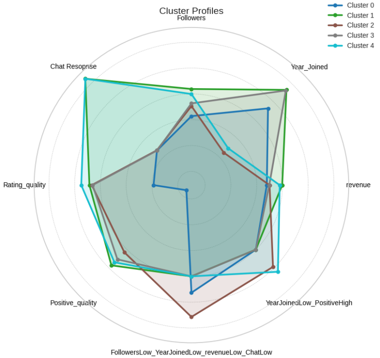

Figure: Cluster Analysis Visualization

4. CONCLUSION

Key Findings:

- FP-Max + clustering improves metrics (Silhouette, Davies-Bouldin, Calinski-Harabasz).
- Uncovers hidden data relationships, enhancing store segmentation.

Impact:

- Practical value for Tiki and similar platforms in competitive e-commerce markets.
- Supports tailored business strategies (e.g., marketing, product optimization).

Future Directions:

- Apply to larger, complex datasets across platforms.

Thank You
Questions & Discussions