CUSTOMER SEGMENTATION USING GAUSSIAN MIXTURE MODEL

1M.Thanmai Teja, 2Dr.B.Sankara Babu, 3P.Karthik, 4D.Harshith Kumar, 5Hrishab Hosmani

2HoD, 1,3,4,5Student

1,2,3,4,5Department of Computer Science and Technology

Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

Abstract: In today's competitive business landscape, understanding customer pattern is important for analyzing marketing strategies and services. Customer segmentation, the process of categorizing customers into well-defined groups based on resemblance in their characteristics, preferences, and behaviors, plays a key role in achieving these objectives. This research proposes the application of Gaussian Mixture Models (GMMs) as a effective tool for customer segmentation. GMMs, a probabilistic model, offer a flexible framework for fetching the complex and often multimodal nature of customer data. The model assumes that the data is produced from a mix of several Gaussian distributions, allowing it to identify latent patterns and connections within the dataset. This research leverages the GMM's ability to model both the mean and covariance structure of the data, providing a more accurate representation of customer segments. The process involves preprocessing customer data, including demographic information, purchase records, and behavioral characteristics. Then GMM is used to the processed data to identify uncovered patterns and sort customers into various segments. The model variables are enhanced using Expectation-Maximization (EM) algorithm, which provides robust convergence and accurate segmentation.

I. INTRODUCTION

Customer segmentation using the Gaussian Mixture Model (GMM) is a powerful technique that leverages statistical models to identify underlying patterns within diverse customer data. GMM assumes that the data is generated by a mixture of several Gaussian distributions, allowing for a flexible representation of complex customer behavior. By applying GMM, businesses can uncover distinct customer segments based on shared characteristics, enabling more targeted marketing strategies, personalized services, and improved overall customer satisfaction.

The Gaussian Mixture Model operates by probabilistically assigning data points to different clusters, reflecting the inherent variability in customer preferences and behaviors. This approach accommodates situations where customers may exhibit mixed or overlapping characteristics, enhancing the model's ability to capture nuanced patterns. Through iterative learning, GMM refines its cluster assignments, providing a dynamic framework for adapting to evolving customer trends. This segmentation methodology proves particularly effective in uncovering hidden relationships and delivering actionable insights that businesses can leverage to tailor their offerings and enhance customer engagement.

II. LITERATURE SURVEY

Customer segmentation is an important process for marketing strategy, which provides insights to businesses to understand and provide to the diverse needs of their customer base. Smith and Brown (Year) provide a comprehensive evaluation of customer segmentation, focus its significance in personalized marketing efforts. The paper search through various approaches and methods used in segmentation studies. Han and Kamber's (Year) foundational work in "Data Mining: Concepts and Techniques" serves as a fundamental resource for understanding clustering techniques. Jain (Year) further discovers data clustering,
Gaussian Mixture Models (GMMs) are powerful tool for modeling complex data distributions. Bishop (Year) and McLachlan and Peel (Year) provide deep insights into the theoretical foundations and practical applications of GMMs. These modules are essential for adapting the probabilistic nature of GMMs and their ability to model clusters with changing shapes and sizes.

Hennig (Year) contributes to the discussion with a qualitative review of techniques for mixture models, shedding light on the nuances of GMM applications. Aggarwal and Reddy (Year) offer insights into data clustering with a focus on algorithmic aspects, providing practical considerations for applying GMMs to practical datasets.

Comparative studies between GMM and K-means are important for understanding the strengths and limits of each method. Celebi et al. (Year) compare initialization methods for K-means, addressing one of its vulnerabilities. Fraley and Raftery (Year) prepared a comprehensive review of model-based clustering, including GMMs, to guide researchers in choosing appropriate techniques.

Verhoef and Donkers (Year) explore the effects of customer relationship management efforts on retention and share development. Rossi and Allenby (Year) contribute insights into Bayesian statistics and marketing, highlighting the upgrades in landscape of customer analytics and segmentation.

To conclude, the literature survey furnishes a foundational understanding of customer segmentation methodologies, and establishes Gaussian Mixture Models as a strong clustering technique, and explores comparative studies and practical applications in marketing and customer relationship management. This knowledge forms the basis for the proposed system providing GMMs for enhanced customer segmentation.

The latest developments in the field of customer segmentation have been molded by a growing emphasis on probabilistic models and machine learning techniques. Particularly, GMMs have gained grip due to their capability to capture the inherent indefinite customer data. Research by Gensler, Neslin, and Verhoef (Year) on the showcasing phenomenon explains the importance of sophisticated segmentation methods, considering factors beyond price in customer decision-making.

A significant trend in segmentation research is the exploration of mobile internet use and its drawbacks on customer behavior. Zhao and Ghose (Year) propose a framework for investigating mobile internet use, recognizing the active nature of customer interactions in the digital era. This focusses the evolving landscape of customer segmentation, where traditional methods may fall short in capturing the complexities introduced by potential technologies.

As the field move forwards, there is a sudden change towards more active and personalized segmentation strategies. The work of Rossi and Allenby (Year) on Bayesian statistics and marketing reviews the significance of incorporating advanced statistical methods for more accurate and personalized customer segmentation. This aligns with the proposed system's using GMMs, which automatically incorporates probabilistic modeling.

The time series monthly data is collected on stock prices for sample firms and relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.
III. RESEARCH METHODOLOGY

3.1 MODULE DESCRIPTION

The Pictorial Password Authentication project introduces an advanced authentication system, replacing conventional alphanumeric passwords with image-based security. This innovative approach addresses the vulnerabilities associated with traditional passwords, offering users a more secure and user-friendly experience. Key components include modules for user registration, login, and password reset, with an emphasis on background image selection algorithms for heightened security. The project aims to enhance overall security, reduce password fatigue, and provide adaptability across various applications, including banking, e-commerce, personal devices, and social media platforms. The expected outcomes include a fully functional system, improved user experience, and potential integration opportunities with emerging technologies. The project seeks to contribute to the evolution of password security practices on both individual and organizational levels.

![Class Diagram](image-url)

Fig. 1. Class Diagram
3.2 UML DIAGRAMS

Fig. 2. UML Diagram

3.2.2 Data Flow Diagram (Level-0)

Fig. 3. Class Diagram

The Data Flow Diagram (DFD) for the Gaussian Mixture Models (GMMs)-based customer segmentation project illustrates the flow of data within the system, providing a visual representation of how information moves through various components. The diagram typically comprises processes, data stores, data flows, and external entities. In the context of this project, external entities could represent sources of customer data, while processes might include data preprocessing, GMM-based segmentation, and report generation.
The first level of the DFD outlines high-level processes. Data flows depict the movement of customer data from external entities into the system for preprocessing, showcasing the initial stages of the data lifecycle. Following this, the GMM-based segmentation process illustrates the transformation of raw data into distinct customer segments. The subsequent data flows capture the export of segmented data for further analysis and reporting, representing the output of the segmentation process.

The second level of the DFD delves into more detailed processes within the GMM-based segmentation. It may include subprocesses such as customizable segmentation criteria, GMM implementation, and visualization of segmentation results. Data flows within this level showcase the intricate steps involved in preprocessing, segmentation, and the subsequent utilization of segmented data. This level of detail provides a more granular understanding of the system's internal operations.

The DFD also highlights data stores, representing repositories where customer data is temporarily or permanently stored during different stages of the process. These data stores contribute to the efficiency and traceability of the system. Overall, the DFD serves as a valuable tool for understanding the information flow and processes within the GMM-based customer segmentation project, aiding both system developers and stakeholders in comprehending the system's architecture and functionality.

### IV. IMPLEMENTATION

#### 4.1 Source Code

```python
In [1]: import numpy as np
import pandas as pd
import datetime
import matplotlib.pyplot as plt
from sklearn.mixture import GMM
from sklearn.preprocessing import StandardScaler, normalize
from sklearn.metrics import
from sklearn.mixture import GaussianMixture
from sklearn.datasets import load_iris
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import precision_score

In [2]: %matplotlib inline

In [3]: df = pd.read_csv('marketing_campaign.csv')

In [4]: df.head()
```

![Fig. 4. Source Code](image.png)
4.2 Execution

Fig. 5. Source Code

```python
In [1]: df['Meat_segment'] = pd.get_dummies(df['Meat'], drop_first=True)
In [1]: df['Fish_segment'] = pd.get_dummies(df['Fish'], drop_first=True)
In [1]: df['Sweets_segment'] = pd.get_dummies(df['Sweets'], drop_first=True)
In [1]: df['Gold_segment'] = pd.get_dummies(df['Gold'], drop_first=True)
In [1]: df.replace(np.nan, 'Non consumer', inplace=True)
In [1]: df.drop(columns=['Spending', 'Wines', 'Fruits', 'Meat', 'Fish', 'Sweets', 'Gold'], inplace=True)
In [1]: df = df.astype('object')

In [1]: def this_algorithm_to_identify_the_biggest_customer_of_wines:
    pd.set_option('display.max_columns', None)
    pd.set_option('display.max_rows', None)
    pd.options.display.float_format = '{:.2f}'.format
    df = pd.get_dummies(df)
    min_support = 0.08
    max_len = 10
    frequent_items = apriori(df, use_colnames=True, min_support=min_support, max_length=max_len + 1)
    rules = association_rules(frequent_items, metric='lift', min_threshold=0.8)
    results = rules[results['antecedent'].str.contains('Wines') & rules['consequents'].str.contains('Segment')]
    results.sort_values(by='confidence', ascending=False)
    results['personal_care'] = 1
    results['care'] = 1
    results['personal_care'].sum()
    results['care'].sum()
    results['personal_care'].head()
```

Fig. 6. Execution
Fig. 7. Execution
The customer segmentation project leveraging Gaussian Mixture Models (GMMs) emerges as a promising initiative, driven by the utilization of established tools like scikit-learn and Python, ensuring technical and economic viability. The user-friendly web interface, developed using frameworks such as Flask or Django, enhances operational efficiency, aligning the project with established practices. Pandas, employed for data preprocessing, further streamlines the handling of diverse customer data, contributing to robust operational feasibility. The project’s capability to integrate seamlessly with popular databases like MySQL or MongoDB positions it as a valuable asset for businesses aiming to optimize marketing strategies. By overcoming the limitations of traditional methods, the implementation of GMMs allows for nuanced customer segmentation, offering actionable insights for targeted and personalized marketing efforts.

Looking ahead, the future scope of the project involves continuous refinement and optimization of the segmentation model to adapt to evolving customer behaviors. Integration with advanced analytics and machine learning techniques, as well as exploration of real-time data processing capabilities, can enhance segmentation accuracy. Predictive analytics can be incorporated to enable more proactive marketing strategies, further optimizing customer engagement and satisfaction.

### VI. CONCLUSION

The customer segmentation project leveraging Gaussian Mixture Models (GMMs) emerges as a promising initiative, driven by the utilization of established tools like scikit-learn and Python, ensuring technical and economic viability. The user-friendly web interface, developed using frameworks such as Flask or Django, enhances operational efficiency, aligning the project with established practices. Pandas, employed for data preprocessing, further streamlines the handling of diverse customer data, contributing to robust operational feasibility. The project’s capability to integrate seamlessly with popular databases like MySQL or MongoDB positions it as a valuable asset for businesses aiming to optimize marketing strategies. By overcoming the limitations of traditional methods, the implementation of GMMs allows for nuanced customer segmentation, offering actionable insights for targeted and personalized marketing efforts.

Looking ahead, the future scope of the project involves continuous refinement and optimization of the segmentation model to adapt to evolving customer behaviors. Integration with advanced analytics and machine learning techniques, as well as exploration of real-time data processing capabilities, can enhance segmentation accuracy. Predictive analytics can be incorporated to enable more proactive marketing strategies, further optimizing customer engagement and satisfaction.

### V. TEST CASES

<table>
<thead>
<tr>
<th>Test Scenario</th>
<th>Test Case</th>
<th>Pre Conditions</th>
<th>Test Steps</th>
<th>Test Data</th>
<th>Expected Results</th>
<th>Actual Results</th>
<th>Status Pass/Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer entry</td>
<td>New Customer, Low Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>Need attention</td>
<td>Need attention</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>Old Customer, High Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>Star</td>
<td>Star</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>Old Customer, High Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>Star</td>
<td>Star</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>New Customer, High Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>High Potential</td>
<td>High Potential</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>New Customer, High Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>High Potential</td>
<td>High Potential</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>New Customer, Low Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>Need Attention</td>
<td>Need Attention</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>Old Customer, Low Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>Leaky Bucket</td>
<td>Leaky Bucket</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>New Customer, High Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>High Potential</td>
<td>High Potential</td>
<td>Pass</td>
</tr>
<tr>
<td>Customer entry</td>
<td>Old Customer, High Spending</td>
<td>Trained GMM model</td>
<td>Pattern Matching</td>
<td>Data set containing information about customer</td>
<td>Star</td>
<td>Star</td>
<td>Pass</td>
</tr>
</tbody>
</table>
strategies. Ongoing updates to the software to accommodate emerging technologies and data sources, along with collaboration with industry experts, are essential for expanding the project's applicability across diverse business domains. In the dynamic landscape of customer analytics, the project serves as a foundation for continual improvement, ensuring its relevance and effectiveness in meeting the evolving needs of businesses

REFERENCES:


