Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Market Segmentation Through Finite Mixture Regression Models with Generalized Normal Distribution and Hierarchical Clustering

S A V S Sambha Murthy S [1], Kunjam Nageswar Rao [2] K. Srinivasa Rao [3]

[1] Research Scholar, Department of Computer Science & Systems Engineering, College of Engineering (A), Andhra University, Visakhapatnam, Andhra Pradesh, India.
sivakotimurthy45@gmail.com.

[2] Professor, Department of Computer Science & Systems Engineering, College of Engineering (A), Andhra University, Visakhapatnam, Andhra Pradesh, India.
kunjamnag@gmail.com

[3] Senior Professor, Andhra University, Visakhapatnam, Andhra Pradesh, India.
ksraoau@yahoo.co.in

ARTICLEINFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Market segmentation through mixture regression models received lot of impetus to its ready applicability in market analytics, business analytics, financial analytics, supply chain analytics , Human Resource analytics etc. In regression analysis it is common to assume that error term follows a normal distribution. Normal distribution has several drawbacks such as being mesokurtic and the model may not fit well for all types of data. Hence, in this paper we develop market segmentation method though mixture of regression models with Generalized Normal Distributed (GND) errors. The GND includes leptokurtic, platykurtic and normal distribution as special cases. The parameters of the proposed model are estimated using Expectation Maximization (EM) algorithm. The initialization of the model parameters is done by using hierarchical clustering algorithm. The segmentation algorithm is obtained through component maximum likelihood under Bayesian framework. The applicability of the proposed algorithm is demonstrated with market segmentation data. The performance of the algorithm is evaluated by computing segmentation performance metrics. It is observed that this method performs much better than the earlier segmentation methods having normal distributed and generalized normal distributed errors with k-means algorithm for the data sets having leptokurtic and platykurtic response variables.

Keywords: Segmentation Methods, Generalized Normal Distribution, Market Segmentation, Regression Analysis, Hierarchical Clustering.

1. INTRODUCTION

Segmentwise Linear Regression (SLR) is a statistical technique that addresses a fundamental limitation of traditional linear regression, which is the inability to capture complex relationships within a dataset that contains different subgroups [1]. In the literature it is also referred to as regression clustering, switching regression. The aim of SLR is to find a given number of linear functions each approximating a subset of the whole data set by minimizing the overall sum of regression errors. SLR can be considered as extension of linear regression. One linear function is used to fit the whole data set in the linear regression where as SLR approximates the data using more than one linear functions. SLR has been applied to several application domains including customer benefit segmentation [10], market segmentation[11], modeling of the metal inert gas welding process [12] , pavement management systems[13], rain fall prediction[14]and PM10 prediction[15].

Wayne S. Desarbo et.al [2] presented a conditional mixture, maximum likelihood methodology for performing clusterwise linear regression. This methodology estimates separate regression functions and membership in S segments or groups simultaneously. Qiang Long et.al [3] described various methods to solve clusterwise linear regression problems. Ye Chow Kuang et.al [1] presented the performance characterization of clusterwise linear regression algorithms. Yifan Zhang et.al [16] studied generalized ordinal Bayesian finite mixture regression model for market segmentation which allows simultaneous variable selction within each derived segment and recovers segment profiling using concomintant variables. Ting Li et. Al [18] extended the classical clusterwise linear regression to incorporate multiple functional predictors by representing the functional coefficients in terms of a functional principal component basis. Kaisa Joki et. Al [19] studied a model and solved the SLR problem by using support vector machines for regression to approximate each cluster. Paul W. Murray et. Al [20] applied data mining methods to identify behavior patterns in historical noisy delivery data in market segmentation. Cathy W.S. Chen et. Al [21] studied a Bayesian approach to simultaneously classify observations drawn from a finite mixture and estimate regression model parameters.

The concept of market segmentation emerged in marketing. Market Segmentation is defined as representing a heterogeneous market as a set of homogeneous submarkets. Segmentation involves creating groups of customers who show similar characteristics and can be targeted with customized strategies in context of product markets. Market segmentation is the process of segmenting a market into distinct groups of customers who share similar characteristics, needs, or behaviors. This approach enables the companies to customize their business strategies for each segment to improve the company sales and profits. [24]. Philippe Masset[23] applied market segmentation to wine data to predict the price of fine wines over their life cycle using regression approach. Tuma, M et.al[25] reviewed finite mixture models in market segmentation. Juan Prieto-Rodriguez et.al [26] investigated whether the null hypothesis of a unique segment of prices in the high end of art market can be rejected using Finite Mixture Model(FMM). Aytaç B et.al [27] studied two regression-based techniques used to detect herding among investors. Herding is described as the tendency of investors to imitate others by suppressing their own beliefs. They also introduced an approach based on the autocorrelation of returns and tested all models on a unique dataset of wine prices. Renneboog L et.al[28] examined geographical segmentation and its effects on price formation and returns in the international art auction market. Arouri M.E et.al [29] presented a theoretical Capital Asset Pricing Model (CAPM) to price assets in different market structures and analyzed whether when markets are partially segmented using the local or the global CAPM yields significant errors in the estimation of the cost of capital for a sample of firms from developed and emerging countries. Ashish Sood et.al[30] studied a model for predicting market penetration of new products through functional regression. Carsten Hahn et.al [31] developed an approach for capturing unobserved customer heterogeneity in structural equation modeling by using a modified finite-mixture distribution approach based on partial least squares. Clusterwise linear regression models are used to build efficient strategic decision making models in the field of market analytics.

In all these papers, it was assumed that attributes of the segmentation data set follows normal distribution and the whole data set is represented by mixture of normal distributions. The major drawback of the Normal mixture model is it assumes attribute vector is mesokurtic. In some data sets the attribute vector associated with data may not have mesokurtic distribution. Hence, to build accurate modeling, it is necessary to generalize the normal mixture model. One of the generalization is including platy, lepty and meso kuritc distributions. Generalized normal distribution is capable of describing platy, lepty and meso kurtic distributions. Very little work has been reported in the literature regarding market segmentation using mixture regression models with Generalized Normal Distributed errors. To develop efficient market segmentation , in this paper an algorithm is developed assuming that the attribute vector associated with the data set follows a generalized normal mixture model proposed method is applied super market dataset[22] to divide customers into low profit, medium profit and high profit margin contributed customers based on their category to the super market store.

The rest of the paper is presented as follows. Section 2 is concerned with Mixture of regression models with Generalized Normal Mixture Model. Section 3 provides the hierarchical clustering algorithm for identifying the number of clusters in the data. Section 4 deals with the initialization of the model parameters. Section 5 describes the estimation of model parameters using Expectation and Maximization (EM) algorithm. Section 6 elaborates the segmentation algorithm for regression models with Generalized Normal Mixture model. Section 7 deals with experimental results and performance evaluation of the model. Section 8 deals with conclusions.

In this paper we follow the following notations:

$i = 1,2,3, \dots \dots . N$ are subjects / observations/ data points.

$j = 1,2,3, \dots \dots . . M$ are regressor variables/attributes.

B_i = the value of the response variable for data point

a_{ij} = the value of the j^th regressor variable.

s= 1,2, … ….S segments

2) FINITE MIXTURE OF REGRESSION MODELS WITH GENERALIZED NORMAL DISTRIBUTION

The finite mixture of regression models are composed through the conditional mixture and maximum likelihood methodology. The SLR models based on the maximum likelihood methodology are also called as finite mixture models for regression problems [7] and finite mixture of linear regression [8]. Finite mixture models for regression were discussed in [9]. In the 1990s, these models were extended by mixing standard linear regression models and generalized linear models [10].

In the finite mixture model method, it is assumed that the observations arise from s distinct random segments [2]. Each of the segments is modeled by specific probability density function. Let w be a random variable and P(w, θk)be a probability density function for each segment s=1,2,.. S . Then the variable w is said to arise from a finite mixture model if it has a density function in the following represented form.

$$h(w, \varphi) = \sum_{s=1}^{S} \alpha_k \, P(w, \varphi s) \quad \alpha_s \geq 0, \sum_{s=1}^{S} \alpha_s = 1 \qquad (2.1)$$

where φs is the parameter vector of the segment for the density function and α_s is the mixing proportion of the segment s, s= 1,2, …. S.

The density function P can be used to formulate the relationship between the regressor and response variables in the regression. Let A is independent attribute vector, B is response attribute vector of a dataset D and assume that B is distributed as a finite mixture of conditional Generalized Normal densities.

The probability density function (pdf) of the Generalized Normal Distribution (GND) with mean$\mu = 0$ is defined as

$$f(x, \theta) = \frac{\theta k(\theta)}{2\sigma} e^{-A(\theta)\left|\frac{x}{\sigma}\right|^{\theta}} \qquad (2.2)$$

Where $A(\theta) = \left(\frac{\Gamma\left(\frac{3}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)}\right)^{\frac{\theta}{2}}, k(\theta) = \frac{\Gamma\left(\frac{3}{\theta}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{\theta}\right)^{\frac{3}{2}}}$, σ is standard deviation, θ is shape parameter and $\Gamma(.)$ is Gamma function.

Figure 2.1 represents the frequency curve of Generalized Normal Distribution with different shape parameters

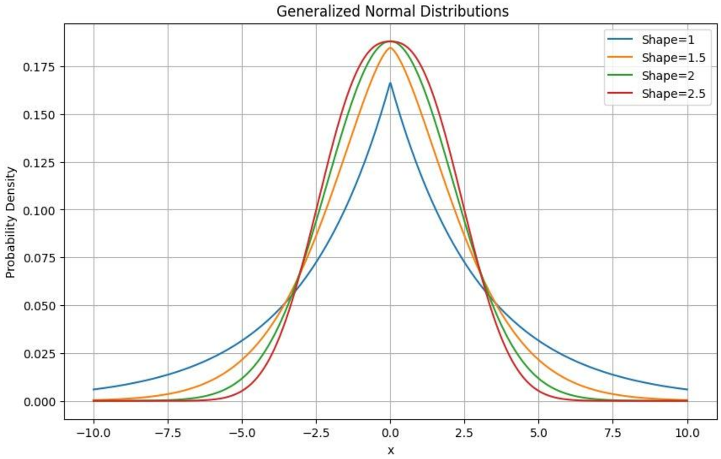

Figure 2.1

The finite mixture regression model of s components is

$$h(B|A, \varphi) = \sum_{s=1}^{S} \alpha_k \, P(B|A, \varphi s) \quad \alpha_s \geq 0, \sum_{s=1}^{S} \alpha_s = 1 \qquad (2.3)$$

where, $P(B|A,\varphi s)$ is the probability density function of the s^th^ component and φ is the vector of all parameters. Then SLR is modeled as a finite mixture or sum of conditional univariate densities as

$$B_i \sim \sum_{s=1}^{S} \alpha_s \, Pij(Bi|A,\varphi s) \qquad (2.4)$$

Where, P_{ij} are univariate Generalized Normal densities. The model becomes a mixture of standard linear regression models. If P_{ij} are members of the exponential family then we get a mixture of generalized linear regression models[11].

A mixture model based approach to regression analysis assumes that the observations of a data set originate from various segments with unknown segment affiliation.

The mixture of linear regression is defined as follows.

$$B_i = \sum_{s=1}^{S} \alpha_s f_s(B_i|s) + \epsilon \quad where \ i = 1,2,3, \dots . I \qquad (2.5)$$

B_i is the dependent variable , α_s is the relative size (mixture proportion) of segment s. where $\sum_{s=1}^{S} \alpha_s = 1$ and $\alpha_s > 0 \ \forall \ s = 1,2, \dots \dots . S$

Now B_i is distributed as a finite sum or mixture of conditional univariate Generalized Normal Distribution (GND).

$B_i = \sum_{s=1}^{S} \alpha_s f_{is}\left(b_i \middle| a_{ij}, \sigma, \beta_{ij}\right)$ Whereβ_{ij} is regression coefficient.

$$B_i = \sum_{s=1}^{S} \alpha_s \frac{\theta_s k(\theta_s)}{2\sigma_s} e^{-A(\theta_k)\left(\frac{|B_i-(\beta_0+\beta_{1s}a_{1i}+\beta_{2s}a_{2i}+ \dots\dots +\beta_{ns}a_{ni})|}{\sigma_k}\right)^{\theta_s}} \qquad (2.6)$$

3. HIERARCHICAL CLUSTERING ALGORITHM FOR IDENTIFYING THE NUMBER OF CLUSTERS UNDER REGRESSION ANALYSIS

In order to utilize the EM algorithm we have to initialize the model parameters which are generally considered as known apriori. The following steps involved in the hierarchical clustering algorithm [6].

Step 1: Start by assigning each observation to a segment. Each of the N observations, are associated with N segments, each containing just one item. Let the distances (similarities) between the segments be the same as the distances (similarities) between the items they contain.

Step 2: Find the most similar pair of segments and merge them into a single segment. The number of segments is now reduced by one. Compute distances (similarities) between the new segments and each of the old segments.

Step 3: Repeat steps 2 and 3 until all items are segmented.

Step 3 can be done in different ways, namely a) Single-Linkage b) Complete-Linkage and c) Average- Linkage segmenting.

4. INITIALIZATION OF MODEL PARAMETERS

The process of identifying the initial estimates of the parametric set for the given linear regression model based on GND, one need to update the parameters using EM algorithm. The main constraint in the execution of EM algorithm is that it is totally dependent on the number of clusters and initial estimates of the model parameters [5]. The initial estimates of the parameters in regression analysis are estimated using moment method of estimation and ordinary least square method of estimation.

The updated equations are to be calculated for α_s (the mixing parameter), σ_s(Standard Deviation) and β_{js} (Regression Coefficient). Since the process is unsupervised, the initial knowledge about the parameters within the data is highly unpredictable.

In order to estimate the values, the methodology of likelihood estimates is subjected.

5. ESTIMATION OF THE MODEL PARAMETERS USING EM ALGORITHM:

In this section, estimation of model parameters using Expectation Maximization (EM) algorithm that maximizes the likelihood function of the model are consider [4]. Given a sample of N observations we can form the likelihood function

$$L = \prod_{i=1}^{N} \left[\sum_{s=1}^{S} \alpha_s \frac{\theta_s k(\theta_s)}{2\sigma_s} e^{-A(s)\left(\frac{|B_i-(\beta_0+\beta_{1s}a_{1i}+\beta_{2s}a_{2i}+\,\ldots\ldots+\beta_{ns}s_{ni})|}{\sigma_k}\right)^{\theta_s}} \right] \quad (5.1)$$

$$\text{where } 0 \le \alpha_s \le 1, \sum_{s=1}^{S} \alpha_s = 1, \sigma_s > 0$$

The log likelihood function is

$$\ln L = \sum_{i=1}^{N} ln \left[\sum_{s=1}^{S} \alpha_s \frac{\theta_s k(\theta_s)}{2\sigma_s} e^{-A(s)\left(\frac{|B_i-(\beta_0+\beta_{1s}a_{1i}+\beta_{2s}a_{2i}+\,\ldots\ldots+\beta_{ns}s_{ni})|}{\sigma_k}\right)^{\theta_s}} \right] \quad (5.2)$$

To estimate the values of parameters $\alpha_s, \sigma_s, \beta_{js}$, EM algorithm consists of two steps i.e. Expectation (E) step and Maximization (M) step is applied . The basic step in the EM algorithm needs the estimation of initial estimates from a given dataset. The final estimates of parameters $\alpha_s, \sigma_s, \beta_{js}$are obtained by maximizing the expected value likelihood or log likelihood. The procedure given by [5] is used to estimate the shape parameter θ_s.

The idea of the EM algorithm is then to iteratively calculate the maximum likelihood estimate of the unknown parameter set $\varphi = (\alpha_s, \sigma_s, \beta_{js})$. The first step of EM algorithm is to estimate initial model parameters $\alpha_s, \sigma_s, \beta_{js}$ from a given observations of data. The second step is to maximize Q(φ, $\varphi^{(1)}$) [6]. Using the steps in the EM algorithm, we get the following updated equations for the model parameters.

for α_s :

$$\alpha_s = \frac{\sum_{i=1}^{S} \widehat{p_{is}}}{I} \quad (5.3)$$

forβ_{js}:

$$\sum_{i=1}^{N} \widehat{p_{is}} \frac{A(\theta_s)}{{\sigma_s}^{\theta_s}} \theta_s \, sgn\left(B_i - \left(\beta_0 + \sum_{i=1,j=1,s=1}^{i=N,j=M,s=S} \beta_{js} a_{ij}\right)\right) \left|B_i - \left(\beta_0 + \sum_{i=1,j=1,s=1}^{i=N,j=M,s=S} \beta_{js} x_{ij}\right)\right|^{\theta_s - 1} x_{ij} = 0 \quad (5.4)$$

As a special case if $\theta_s = 2$ we have normal distribution. Then for $\theta_s = 2$ we have

$$\sum_{i=1}^{N} \widehat{p_{is}} \left| B_i - \left(\beta_0 + \sum_{i=1,j=1,s=1}^{i=N,j=M,s=S} \beta_{js} x_{ij} \right) \right| x_{ij} = 0$$

for σ_s: $\sum_{i=1}^{N} \frac{\widehat{p_{is}}}{\sigma_s} \left(\theta_s A(\theta_s) \left|B_i - \left(\beta_0 + \sum_{i=1,j=1,s=1}^{i=N,j=M,s=S} \beta_{js} x_{ij}\right)\right|^{\theta_s} {\sigma_s}^{-\theta_s} - 1\right) = 0$ (5.5)

As a special case if $\theta_s = 2$ we have normal distribution. Then for $\theta_s = 2$ we have

$$\sigma_s = \left(\frac{\sum_{i=1}^{N} \widehat{p_{is}} \left(\left| B_i - \left(\beta_0 + \sum_{i=1,j=1,s=1}^{i=N,j=M,s=S} \beta_{js} x_{ij}\right) \right| \right)^2}{\sum_{i=1}^{I} \widehat{p_{is}}} \right)^{\frac{1}{2}}$$

Solving the equations (4.3),(4.4) and (4.5) simultaneously and iteratively , the refined estimates of the model parameters $\alpha_s, \sigma_s, \beta_{js}$ can be obtained.

Once estimates of $\alpha_s, \sigma_s, \beta_{js}$ are obtained, one can assign each observation i to each segment s though the estimated posterior probability using Bayes rule.

$$\widehat{p_{is}} = \frac{\widehat{\alpha_s} f_{is}\left(B_i \middle| x_{ij}, \widehat{\alpha_s}, \widehat{\beta_{is}}\right)}{\sum_{s=1}^{S} \widehat{\alpha_s} f_{is}\left(B_i \middle| x_{ij}, \widehat{\alpha_s}, \widehat{\beta_{is}}\right)} \quad (5.6)$$

Assign observation i to segment s iff $\widehat{p_{is}} > \widehat{p_{il}} \forall\, l \neq s = 1,2, \ldots \ldots S$.

EXPECTATION – MAXIMIZATION ALGORITHM

Step1: Select the initial model parameters.

Step 2: obtain revised estimates of the model parameters $\alpha_s, \sigma_s, \beta_{js}$ using equations (5.3), (5.4) and (5.5)

Step 3: Repeat the process until the parameters do not change or the difference in successive computations is within the given threshold value.

Step 4 : Write the final estimates of parameters α_s , σ_s, β_{js}

6. SEGMENTATION ALGORITHM FOR REGRESSION WITH GENERALIZED NORMAL MIXTURE MODEL

In this section, the segmentation algorithm for regression models with Generalized Normal Distribution is presented for identifying the new observations with one of the available clusters. The steps involved in this algorithm are as follows.

Step 1: Draw the dendogram for the training data set in order to obtain the initial number of clusters by using the hierarchical clustering algorithm.

Step 2: Obtain initial estimates of the model parameters.

Step 3: Obtain the final estimates of the model parameters using the updated equations of the EM algorithm given in section 4.

Step 4: For a new observation, compute the conditional likelihood with the model parameters of the s^{th} class and assign it to the class for which the sample conditional likelihood is maximum. i.e. the classification is $C=\text{argmax}_k P(D_t | C_k)$. Where C is the maximum likelihood class and D_t is the new observation.

7. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

In this section, the applicability of the proposed algorithm for segmenting marketing data is demonstrated. The dataset was collected from Kaggle dataset repository [22]. This dataset has 21 features among them Segment (The segment where the Customer belongs), Sales (Sales of the Product), Quantity (Quantity of the Product) and Profit (Profit/Loss incurred) are considered as relevant attributes for this study. Here there are three segments of customer groups like Consumer, Home Office and Corporate. After analyzing super market data set, it was observed that two attributes sales (A_1) and quantity (A_2) are most relevant attributes for deriving the profit(B) margins such as low profit margin, medium profit margin and high profit margin of the store. Here Consumer segment customers contributed low profit margin, Home Office segment customers contributed medium profit margin and Corporate segment customers contributed high profit margin. To identify the margins of the profit, it is required to segment the data set into various clusters based on sales and quantity variables. The number of clusters in the super market data is not known and requires unsupervised learning algorithms to identify various margins of profit. Hence a study is carried out by collecting a sample of 80 data points with sales and quantity variables of super market data set.

Using hierarchical clustering algorithm, the number of profit margins according to sales and quantity is determined. For implementing the hierarchical clustering algorithm, the initial number of clusters is required. Hence, using the training data, the sample observations are plotted in dendogram shown in figure 7.1.

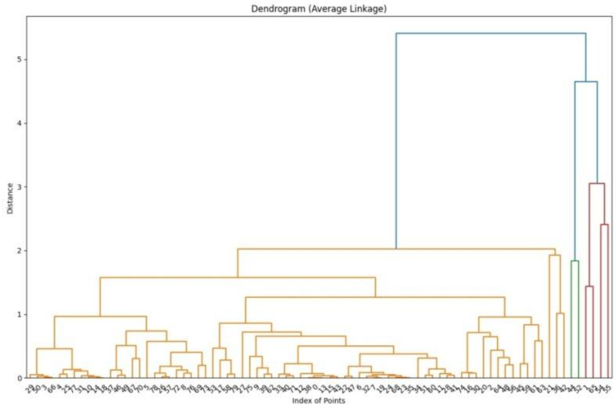

Figure 7.1: Dendogram of Market Data

Using the initialization of parameters discussed in section 6, the initial estimates of parameters α_s , σ_s, β_{js} are obtained for 3 profit margins such that low profit margin corresponds to cluster 1, medium profit margin corresponds to cluster 2 and high profit margin corresponds to cluster 3. The computed initial estimates of the model parameters are presented in Table 7.1

Table 7.1: Initial estimates of the model parameters

Parameter	Segment 1(Low Profit Margin)	Segment 2 (Medium Profit Margin)	Segment 3(High Profit Margin)
α_s	0.1750	0.6500	0.1750
σ_s	117.5365	11.4240	0.0010
β_{js} Intercept	17.9278	-52.9095	0.0010
Coefficient 1	0.3597	0.2356	0.2636
Coefficient 2	-4.6381	-3.3996	-20.1636

Using these initial estimates and the EM algorithm, the refined estimates of parameters for each segment are obtained and presented in Table 7.2.

Table 7.2: Final estimates of the model parameters

Parameter	Segment 1(Low Profit Margin)	Segment 2 (Medium Profit Margin)	Segment 3(High Profit Margin)
α_s	0.5695	0.4055	0.0250
σ_s	4.2448	1.3767 e+03	5.7443e -26
β_{js} Intercept	0.1846	32.4429	2.8715
Coefficient 1	0.3239	0.1414	0.2585
Coefficient 2	0.1034	-6.1929	-20.2312

With these final estimates, the 3 segments of profit margins are estimated as

Segment 1: Consumer segment customers (low profit margin)

$B= 0.1846 + 0.3239 A_1 – 0.1034 A_2$

Here A_1 represents sales, A_2 represents quantity and B represents profit.

Segment 2: Home Office Segment customers (medium profit margin)

$B= 32.4429 + 0.1414 A_1 – 6.1929 A_2$

Segment 3: Corporate Segment customers (high profit margin)

$B= 2.8715 + 0.2585 A_1 – 20.2312 A_2$

Therefore, the model characterizes the whole data set is a three-segment mixture of Generalized normal Mixture Model (GNMM) whose segment proportions are: α_1 =0.5695, α_2=0.4055,α_3=0.0250, respectively. For evaluating the proposed algorithm, the test data consisting of 80 data points is considered. The proposed unsupervised algorithm using GNMM identified 50 tuples as low profit margin, 28 tuples as medium profit margin and 2 tuples as high profit margin. For evaluating the performance of the proposed algorithm, accuracy, misclassification rate, precision, recall and F-measure are used. For the proposed unsupervised learning algorithm of GNMM with hierarchical clustering (GNMM-H), the performance measures for each segment are computed and presented in Table 7.3.

Table 7.3: Performance Measures of the Mixture of GNMM classifier with hierarchical clustering

	True Positive Rate(TPR) Recall	Precision	False Discovery Rate	F-Measure
Segment 1	0.9800	0.9800	0.0200	0.9800
Segment 2	1.0000	0.9285	0.0000	0.9629
Segment 3	0.2500	1.0000	0.5000	0.4000

Table 7.4: Performance Measures of the Mixture of GGMM classifier with k-means clustering

	True Positive Rate(TPR) Recall	Precision	False Discovery Rate	F-Measure
Segment 1	0.9682	0.9839	0.0317	0.9760
Segment 2	0.9166	0.7857	0.0833	0.8461
Segment 3	0.8000	1.0000	0.2000	0.8888

Table 7.5: Performance Measures of the Mixture of GMM classifier

	True Positive Rate(TPR) Recall	Precision	False Discovery Rate	F-Measure
Segment 1	0.9365	0.9672	.0635	0.9516
Segment 2	0.9166	0.7857	0.0833	0.8461
Segment 3	0.8000	1.0000	0.2000	0.8888

To compare the efficiency of the proposed GNMM classifier using hierarchical clustering (GNMM-H) with earlier Generalized Gaussian Mixture Model using k-means (GGMM-K) and Gaussian Mixture Model (GMM) classifiers, recall, precision, false discovery rate and F-measure are computed and presented in Table 7.4 and Table 7.5 respectively.

Comparing Table 7.3, Table 7.4 and Table 7.5 it is observed that the F value for segment 1 and Segment 2 using the proposed classifier is more compared to that of the classifier with GGMM-K and GMM. The f value for segment 3 is less than proposed classifier compare to classifier with GGMM-K and GMM.

To compare the efficiency of the developed unsupervised algorithm with existing unsupervised learning algorithm with GMM model for both sales and quantity variables, the same test data were considered and the accuracy and misclassification rates were computed. Table 7.6 presents the accuracy and error rates of GNMM-H, GGMM-K and GMM classifiers.

Table 7.6: Performance evaluation of accuracy & error rate

Classifier	Accuracy	Error rate
GNMM-H	0.9625	0.0375
GGMM-K	0.9500	0.0500
GMM	0.9250	0.0750

From Table 7.6 it is observed that the accuracy of GNMM-H classifier is more compared to the accuracy of GGMM-K and GMM classifiers and the error rate of GNMM-H classifier is lesser compared to the error rate of GGMM-K and GMM classifiers.

8. CONCLUSIONS

This paper deals with the development and analysis of a new and novel method in segmentation algorithm for market analytics data using mixture regression models with Generalized Normal Distribution with hierarchical clustering. In market segmentation so far the algorithms developed using mixture regression models with normal distribution. For the first time we developed an unsupervised learning algorithm for market segmentation using

Generalized Normal mixture regression models under Bayesian framework. This algorithm is more suitable for analyzing all types of data sets that show different behaviours such mesokurtic, playkurtic and leptokurtic. This algorithm is utilized for analyzing the real time situations in market analytics, business analytics, financial analytics, HR analytics etc. where the variables under study are correlated and follows Generalized Normal Distribution.

Another important feature of this proposed algorithm is integration of hierarchical clustering with model based method in learning algorithms. The learning algorithm is developed based on component maximum likelihood under Bayesian framework. Hence it is assumed that the attribute vector is generated from a heterogeneous population which can be modeled by a finite mixture of regression models with Generalized Normal Distribution. The model parameters are estimated using Expectation and Maximization (EM) algorithm.

The performance of the proposed algorithm is evaluated using the super market data set. The experimental results revealed that the proposed algorithm outperforms the existing learning algorithms. This learning algorithm can be extended to the integration of clustering algorithms with mixture regression models using Truncated Generalized Normal Distributed errors which will be considered later.

REFERENCES

[1] Ye Chow Kuang , Melanie Ooi (2024), Performance Characterization of Clusterwise Linear Regression Algorithms, Wiley Interdicplinary Reviews: Computational Statistics, pp.1-16.

[2] Wayne S. DeSarbo and William L. Cron(1988), A maximum likelihood methodology for clusterwise linear regression. Journal of Classification 5, pp. 249-282

[3] Qiang Long, Adil Bagirov, Sona Taheri , Nargiz Sultanova , and Xue Wu(2023), Methods and Applications of Clusterwise Linear Regression: A Survey and Comparison. ACM Trans.Knowl.Discov. Data. 17, 3,pp 1-54.

[4] Mclanchlan, G. and Peel. D. (2000), The EM Algorithm for parameter estimation, John Wiley and Sons New York.

[5] Shaoquan YU, Anyi Zhang, Hongwei LI (2012) , A Review on estimating the Shape Parameter of Generalized Gaussian Distribution , Journal of Information Systems, 8,21, pp. 9055-9064.

[6] K.Srinivasa Rao et.al(2014) Image Segmentation for Animal Images using Finite Mixture ofPearson type VI Distribution, Global Journal of Computer Science and Technology: F Graphics & Vision, 14,3, pp. 1-13

[7] Bilmes. Jeff A. (1998) A Gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Intl. Computer Science Institute, Berkely.

[8] Susana Faria and Gilda Soromenho(2010). Fitting mixtures of linear regressions. Journal of Statistical Computation and Simulation 80,2 , pp. 201-225.

[9] Richard E. Quandit(1972), A new approach to estimating switching regressions. Journal of American Statistical Association 67, 338, pp.306-310.

[10] Michel Wedel and Wayne S. DeSarbo(1995), A mixture likelihood approach for generalized linear models. Journal of Classification 12, 1, pp. 21-55.

[11] Michel Wedel and Cor Kistemaker (1989). Consumer benefit segmentation using clusterwise Linear regression. International Journal of Research in Marketing 6, 1, pp.45-59.

[12] Christian Preda and Gilbert Saporta(2005). Clusterwise PLS regression on a stochastic process. Computational Statistics & Data Analysis 49, 1, pp. 99-108.

[13] Jagadeesh P. Ganjigatti, Dilip K. Pratihar and A.Roy Choudhury (2007). Global versus cluster-wise regression analyses for prediction of bead geometry in MIG welding process. Journal of Materials Processing Technology 189, 1- 3 , pp. 352-366.

[14] Mukesh Khadka and Alexander Paz.(2017), Comprehensive clusterwise linear regression for Pavement management systems. Journal of Transportation Engineering, Part B : Pavments 143, 4.

[15] Adil M. Bagirov, Arshad Mahmood and Andrew Barton(2017), Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach . Atmospheric Research 188, pp. 20-29.

[16] Jean – Michel Poggi and Bruno Portier(2011) . PM10 forecasting using clusterwise linear regression. Atmospheric Environment 45, 38, pp. 7005-7014.

[17] Yifan Zhang, Ducan K.H. Fong , Wayne S.DeSarbo(2021) , A Generalized Ordinal Finite Mixture Regression Model Market Segmentation, International Journal of Research in Marketing .

[18] Ting Li, Xinyuan Song, Yingying Zhang, Hongtu Zhu, Zhongyi Zhu(2021) , Clusterwise Functional Linear Regression Models , Computational Statistics and Data Analysis , pp. 1-15.

[19] Kaisa Joki , Adil M. Bagirov, Napsu Karmitsa, Marko M. Makela, Sona Taheri (2020) , Clusterwise Support Vector Linear Regression, European Journal of Operational Research , pp. 19-35.

[20] Paul W.Murray, Bruno Agard, Marco A. Barajas (2017), Market Segmentation through data mining: A method to extract behaviors from a noisy data set, Computers & Industrial Engineering, pp. 233-252.

[21] Cathy W.S. Chen, Jennifer S.K. Chan, Mike K.P. So, Kevin K.M. Lee,(2011) Classification in segmented regression problems, Computational Statistics and Data Analysis, pp. 2276-2287.

[22] https://www.kaggle.com/datasets/vivek468/superstore-dataset-final.

[23] Philippe Masset(2024), Market segments and pricing of fine wines over their lifecycle, Economic Modelling,141,106915.

[24] Wedel, M.Kamakura, W.A (2000) , Market Segmentation: Conceptual and Methodological Foundations, Second Edition.

[25] Tuma, M and Decker, R(2013) , Finite Mixture Models in Market Segmentation: A Review and Suggestions for Best Practices, The Electronic Journal of Business Research Methods 11,1, pp 02-15.

[26] Juan Prieto-Rodriguez, Marilena Vecco(2021), Reading between the lines in the art market: Lack of transparency and price heterogeneity as an indicator of multiple equilibria, Economic Modelling, 102, 105587.

[27] Aytaç B, Coqueret G, Mandou C (2018), Herding behavior among wine investors, Economics Modelling, 68, pp.318-328

[28] Renneboog L, Spaenjers C (2014), Investment returns and economic fundamentals in International art markets , in: Canvases and Careers in a Cosmopolitan Culture. On the Globalization of Contemporary Art Markets, O. Velthuis and S. Baia-Curioni (eds.), Oxford University Press.

[29] Arouri M.E, Rault C, Sova R, Sova A(2013), Market Structure and the Cost of Capital, Economics Modelling, 31, pp.664-671.

[30] Ashish Sood, Gareth M. James, Gerard J. Tellis, (2008) Functional Regression: A New Model for Predicting Market Penetration of New Products. Marketing Science 28(1) , pp.36-51.

[31] Carsten Hahn,Michael D. Johnson,Andreas Herrmann, Frank Huber(2002), Capturing Customer Heterogeneity Using A Finite Mixture PLS Approach, Schmalenbach Business Review 54, pp.243 – 269.