RESEARCH PAPER | Statistics | Volume : 4 | Issue : 9 | September 2014 | ISSN - 2249-555X

A Study on Socio-Economic Status of Socially Downtrodden People in the Southern Districts of Tamilnadu Using Fuzzy Stochastic Model

KEYWORDS: Below Poverty Line, Model-based Clustering algorithm, Gaussian Mixture Model, EM-algorithm, Fuzzy logic.

Arumugam.P
Department of Statistics, Manonmaniam Sundaranar Unviersity, Tirunelveli, Tamilnadu

Siva Ambika.S.R
Department of Statistics, Manonmaniam Sundaranar Unviersity, Tirunelveli, Tamilnadu

ABSTRACT *The Socio-Economic status is referred as "use of economics in the study of society". It is a combined measure of person's Income, Health, Education and Community in the society. The planning commission has periodically estimated the poverty line and poverty ratios. According to this, the Below Poverty Line (BPL) data of two districts in south Tamil Nadu are taken in account. In this paper, the model based clustering algorithm is used to classifying and identifying the corresponding components of the observations. The Gaussian Mixture Model (GMM) is fitted to this data, the number of components and the parameters are estimated through EM-algorithm. The fuzzy logic is also used to classifying the observation into linguistic values.*

1. Introduction

Socioeconomic environment refers to a wide range of interrelated and diverse aspects and variables relating to or involving a combination of social and economic factors. These aspects and variables could, in general, be categorized into several categories including, economic, demographic, public services, economic and social. The Socio-economic study is very important to improve the level of living standard of the people. So many Statistical Analyses are applied for the study of Socio-Economic status. In this paper Model based Clustering Algorithm is employed to this study. Generally the clustering algorithms are framed based on distance. The alternative approach for clustering algorithm is probability models, such as the finite mixture model for probability densities which is termed as Model-based clustering algorithm. In this algorithm the data are assumed as that are generated by a mixture of probability distributions in which each component represents a different cluster.

A survey of cluster analysis in a probabilistic and inferential framework was presented by Bock (1996). Early work on model-based clustering can be found in Edwards and cavalla-Sforza (1965), Day (1965), wolfe (1970) and Binder (1978). Some issues in cluster analysis, such as the number of clusters are discussed in Mclachlan and Basford (1988), Banfield and Raftery (1993) Mclachlan and Peel (2000), Everitt et al.,(2001) and Fraley and Raftery (2002). The main objective of this study is to estimate number of clusters and identifying the corresponding clusters of each Panchayats. On comparison of parameters, find how many numbers of Panchayats holds the high number of families under Below Poverty Line.

2. Methodology

2.1 Model Based Clustering Algorithm

Let $D=\{x_1,x_2,\ldots x_n\}$ be a set of observations; let $f_j(x_i|\Theta_j)$ be the density of an observation x_i from the j^{th} component (cluster), where Θ_j are the corresponding parameters and let k be the number of components in the mixture. For instance, assuming the data come from a mixture of Gaussian distributions then the parameters Θ_j consist of a mean vector μ_j and a covariance matrix Σ_j and the density has the form

$$f_j(x_i|\mu_i,\Sigma_j)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}e^{\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)} \quad \ldots (1)$$

Where d is the dimension of the data.

If mixture likelihood approach is used for clustering, it becomes the estimation of the parameters of the assumed mixture model. Mathematically this approach maximizes

$$L_M(\theta_1,\theta_2,\ldots\theta_k,\tau_1,\tau_2,\ldots,\tau_k|D)=\prod_{i=1}^{n}\sum_{j=1}^{k}\tau_j f_j(x_i|\theta_j) \quad \ldots(2)$$

Where $\tau_j \geq 0$ is the probability that an observation belongs to the j^{th} component and $\sum_{j=1}^{k}\tau_j=1$. It can be written in the form $P(x)=\sum_{i=1}^{k}P_i\,p(x|cluster\,i)$, and may assume $P(x|cluster\,i)\sim N(\mu_i,\Sigma_i)$ and $\sum_{i=1}^{k}P_i=1$, $0\leq P_i\leq 1$.

2.2 Parameter Estimation using EM algorithm

The EM algorithm is a general statistical method of maximum likelihood estimation in the presence of incomplete data that can be used for the purpose of clustering. It was first formulated by Dempster et al., (1977).

E- Step

Since one cannot determine which cluster i produce the particular samples (x_i), in the E step one can approximately expect the sample x_i is come from which cluster i.

$$P(cluster\,i|x_j)=\frac{P_iP(x_j|cluster\,i)}{P(x_j)}$$

$$=\frac{P_i\frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}e^{\frac{1}{2}(x_j-\mu_i)^T\Sigma_i^{-1}(x_j-\mu_i)}}{\sum_{k=1}^{K}P_k\frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}e^{\frac{1}{2}(x_j-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)}}$$

… (3)

M-Step

In the M-step one can compute new parameter estimates namely

$$P_i=\frac{1}{n}\sum_{j=1}^{n}P(cluster\,i|x_j) \quad \ldots (4)$$

Volume : 4 | Issue : 9 | September 2014 | ISSN - 2249-555X

$$\mu_i = \sum_{j=1}^{n} x_j \frac{P(cluster\ i|x_j)}{\sum_{j=1}^{n} P(cluster\ i|x_j)}$$

… (5)

$$\Sigma_i = \sum_{j=1}^{n} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \frac{P(cluster\ i|x_j)}{\sum_{k=1}^{K} P(cluster\ k|x_j)}$$

… (6)

By repeating the E-step and the M-step the parameter estimates will converge to the maximum likelihood estimates. The number of clusters and the distribution of the component densities can be considered as producing different statistical models for the data. The final model can be determined by the Bayesian Information Criterion (BIC).

2.3 Fusion of model based clustering algorithm and fuzzy logic

In 1965 Lofti A. Zadeh pioneered fuzzy set theory. Fuzzy set is a set with boundaries that are not precise which are vague and ambiguous. The existence of an element is not a matter of affirmation (non-membership) or denial (full-membership), but rather a matter of degree. The element in a fuzzy set having varying degrees of membership in the set $\underset{\sim}{A}$ $\underset{\sim}{A}$. A membership in a fuzzy set is mathematically represented by membership function $\mu_A(X) = [0,1]$.

The symbol $\mu_A(X)$ is the degree of membership of an element X in a fuzzy set $\underset{\sim}{A}$. Therefore $\mu_A(X)$ is the value on the unit interval that measures the degree to which element X belongs to fuzzy set $\underset{\sim}{A}$. The main idea of fuzzy set theory is linguistic variable, it is fuzzy variable. The subject of the study is linguistic variable and value of the subject is linguistic value. A linguistic variable carries the concept of fuzzy sets quantifies called hedges. The range of possible values of a linguistic variable represents the universe of discourse of that variable. On this basis one can incorporate the fuzzy logic and final model obtained from model-based clustering algorithm. The linguistic values are determined by the parameters of the final model. Then one can find the membership of the each observation in each fuzzy set.

3. Experimental Results

In this paper, for studying the socio-economic level of Tirunelveli and Tuticorin district, the blockwise data are considered. There are 16 and 12 blocks in Tirunelveli and Tuticorin district respectively. These districts hold 354 and 403 Panchayats. For each Panchayats the number of people in the below poverty ling is given in Figure 3.1 and Figure 3.2

After performing the Model-based clustering algorithm to the Panchayats data of two districts, the final model is chosen by Bayesian Information Criterion (BIC). The selection of model (number of clusters) and parameter of distribution by BIC is given in Table 3.1 and Table 3.2.

From the above Table 3.1 and Table 3.2, the Gaussian Mixture Model (GMM) with three components is appropriate for two districts by using EM algorithm for parameter estimation. Since this model has the lowest BIC value. Cluster classification of each district is also estimated by model based clustering algorithm.

Using the estimated parameter of the selected model one can give the linguistic value to the linguistic variable below poverty line, number of component or cluster is 3, so one can create three linguistic values along with this estimated parameters. Cluster classification of each district is also estimated by model based clustering algorithm. It is given in Figure 3.3 and Figure 3.4.

By framing the membership function of the fuzzy set (Below Poverty Line) for this data set, one can easily identify whether a Panchayat has low, high, or medium number of BPL in their Panchayats. For example the Panchayat Pudur in Tirunelveli has the number of BPL is 158. From the membership functions using the maximum rule, it is in the 'low'fuzzy set. The list of Panchayats in Tirunelveli and Tuticorin districts which are all come under the fuzzy set 'high' is given in table 3.3 and table 3.4. The new observation can also easily identified by this membership functions without reconstructing the model.

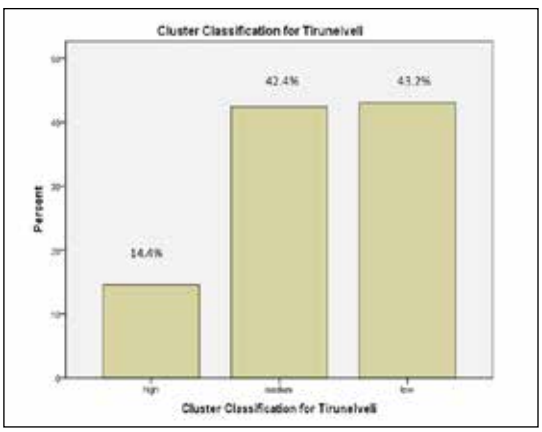

Figure3.3: Cluster classification for Tirunelveli

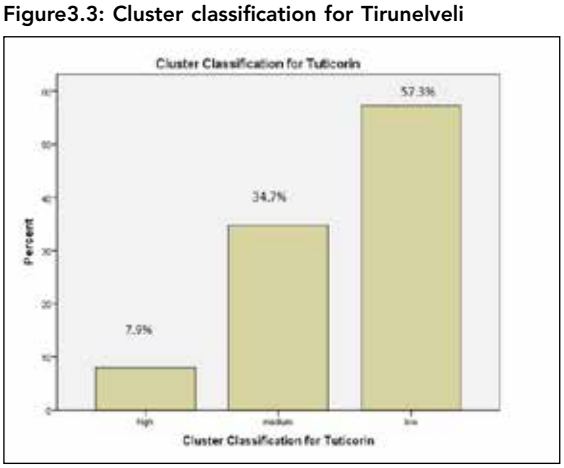

Figure3.4: Cluster classification for Tuticorin District

4. Conclusion

The model based clustering is well performed to identify number of clusters and corresponding clusters of each Panchayats. The membership function is defined using the constructed model. So that cluster of new observation can identify easily. There are 14.4% and 7.9% of Panchayats in the class of high. On comparison of two districts, Tirunelveli districts have more number of Panchayats in the class of high. From the results, the government should give more intension to the Panchayats for which are all comes under the cluster or components or fuzzy set as high for improving the living standard of people.

Acknowledgement

The authors acknowledge university grants commission, New Delhi for providing financial support to carry out this work under UGC's Major Research Project.

Volume : 4 | Issue : 9 | September 2014 | ISSN - 2249-555X

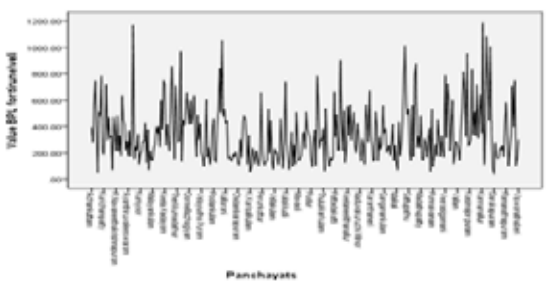

Figure 3.1: The number of people in the Below Poverty Line for each Panchayts in Tirunelveli District

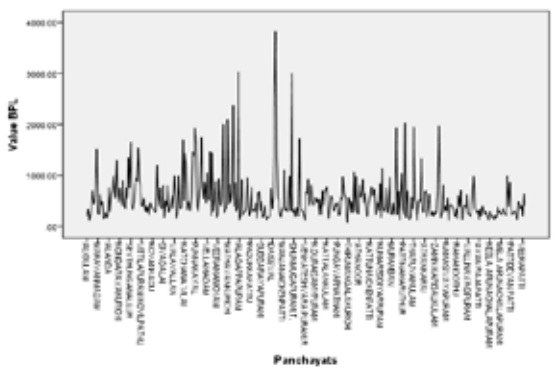

Figure 3.2: The number of people in the Below Poverty Line for each Panchayts in Tuticorin District.

Table3.1: Model selection for Tuticorin District

Number of Components	Components	Mean	Variance	Proportion	Log-likelihood	BIC
1	1	344.6097	$4.6473x10^4$	1	-2384.08	4.779
2	1	218.6258	$8.2897x10^3$	0.561623	-2335.16	4.6996
	2	506.0130	$4.9005x10^4$	0.438377		
3	1	164.2967	$3.1483x10^3$	0.339655	-2320.84	4.6886
	2	345.4864	$1.3187x10^4$	0.451188		
	3	635.5319	$5.1195x10^4$	0.209158		
4	1	166.8343	$3.341x10^3$	0.051467	-2321.11	4.7007
	2	166.8343	$3.341x10^3$	0.211940		
	3	652.5541	$4.9607x10^4$	0.323187		
	4	352.6334	$1.3411x10^4$	0.413406		

Table3.2: Model selection for Tirunelveli District

Number of Components	Components	Mean	Variance	Proportion	Log-likelihood	BIC
1	1	541.9107	20.0921x10+5	1	-3040.43	6.0929
2	1	$1.145x10^3$	$4.3414x10^5$	0.207777	-2877.41	5.7848
	2	506.0130	$4.9005x10^4$	0.792223		
3	1	292.9420	$9.3644x10^3$	0.515878	-2838.86	5.7257
	2	$1.5439x10^3$	$5.0211x10^5$	0.102089		
	3	610.3509	$4.4128x10^4$	0.3820338		
4	1	$1.5608x10^3$	$4.9998x10^5$	0.099665	-2835.25	5.7382
	2	379.0637	$8.7247x10^3$	0.317401		
	3	666.3526	$3.9261x10^4$	0.300655		
	4	232.7340	$4.5502x10^3$	0.282279		

Table 3.3: The list of Panchayats of Tirunelveli in the class of high

Panchayats			
Karungadal	Ottapidaram	Mappillaiurani	Manapadu
Karungulam	Pandavarmangalam	Mela authoor	Manthithoppu
Katchana vilai	Paraman kurichi	Mukkani	Tharuvaikulam
Kattarimangalam	Punnakayal	Nalumavadi	Thittankulam
Kayamozhi	Puthiyamputhur	Inammaniyachi	Vepalodai
Kulasekaran pattinam	Sekkarakudi	Kadalaiyoor	Mookuperi
Kulathur	Seythunganallur	Vallanadu	Iluppaiyoorani
Kurukkuchalai	Srivenkatesapuram	Venkattaramanujapuram	

Table 3.4: The list of Panchayats of Tuticorin in the class of high

Panchayats			
Pappakudi	Aavaraikulam	Therku valliyoor	Sernthamaram Majara
Pappankulam	Ariyanayagipuram	Utthumalai	Servaikaranpatti
Periya Pillai Valasie	Avudiyanoor	Vadakkankulam	Suthamalli
Piranur	Balapathirampuram	Vannikonendai	Thalpathisamudram
Pottalpudur	Chettikulam	Veerasigamani	Thalayuthu
Pudupatti	Devipattanam	Vengadampatti	Thenmalai
Puliyarai	Elathur	Kavalkinaru	Levinjipuram
Punaiyapuram	Erukkandurai	Keela Ambur	Mannarkovil
Ramanathapuram	Gangaikondan	Keela Kadayam	Itamozhi
Kilangadu	Gunaramanallur	Keela veeranam	Kandaganeri
Kutaiyaneri	Kadayam	Kulasekarapatti	Mayamankurichi
Kulasekaramangalam	Kalappakulam	Kuthukalvalasai	Nayinakaram
Karivalamvandanallur	Palankottai	Pallakal	

Volume : 4 | Issue : 9 | September 2014 | ISSN - 2249-555X

REFERENCE [1] Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering.Biometrics, 49(3):803–821. | [2] Binder, D. (1978). Bayesian cluster analysis.Biometrika, 65(1):31–38. | [3] Bock, H. (1996). Probabilistic models in cluster analysis. Computational Statistics and Data Analysis, 23(1):5–28. | [4] Dasgupta, A. and Raftery, A. (1998). Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American Statistical Association, 93(441):294–302. | [5] Day,N. (1969). Estimating the components of a mixture of normal distributions. Biometrika, 56(3):463–474. | [6] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via theEMalgorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38. | [7] Edwards, A. and Cavalli-Sforza, L. (1965). A method for cluster analysis. Biometrics, 21(2):362–375. | [8] Everitt, B., Landau, S., and Leese, M. (2001). Cluster Analysis, 4th edition. New York: Oxford University Press. | [9] Fraley, C. and Raftery,A. (2002). Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association, 97(458):611–631. | [10] Kass, R. and Raftery, A. (1995). Bayes factors. Journal of the American Statistical Association, 90(430):773–795. | [11] McLachlan, G. and Basford, K. (1988). Mixture Models: Inference and Applications to Clustering, volume 84 of statistics: Textbooks and Monographs. New York: Marcel Dekker, Inc. | [12] McLachlan, G. and Peel, D. (2000). Finite Mixture Models. New York: Wiley. | [13] Meng, X. and van Dyk, D. (1997). The EM algorithm—An old folk-song sung to a fast new tune. Journal of the Royal Statistical Society. Series B (Methodological), 59(3):511–567. | [14] Schwarz,G. (1978). Estimating the dimension of a model. The Annals of Statistics, | 6(2), Pp: 461–464. |