Volume-4, Issue-5, October-2014, ISSN No.: 2250-0758
International Journal of Engineering and Management Research
Available at: www.ijemr.net
Page Number: 220-228

Gaussian Mixture Modeling (Gmm) for Cluster Analysis

Srinivas. N[1], Dattatreya Rao.A.V, Karteeka Pavan.K[3]
[1]Andhra Loyola College, Vijayawada, Andhra Pradesh, INDIA
[2]Acharya Nagarjuna University, Guntur, Andhra Pradesh, INDIA
[3]RVR & JC College of Engineering, Guntur, Andhra Pradesh, INDIA

ABSTRACT

Srinivas et al (2014) have suggested Automatic Merging of Clustering (AMOC) as the best optimal clustering algorithm in grouping six datasets viz., Iris, Glass, Breast Cancer, Halfmoon, Path based and Spiral. Each dataset could be made up made up of hidden patterns associated with small subgroups having weak commonality across the whole population forcing us to concentrate model building (a combined representation of unification of entire congregation of a given data set and segregating them into clusters) to have common representation for the newly formed clusters. This paper is aimed at applying finite mixture models (GMMs) for modelling the resultant clusters and to summarize the work done along with conclusions.

I. MODELLING DATASET

In big data era, large sample size enables us to better understand heterogeneity, shedding light towards studies such as exploring association between certain covariates (genetic characteristics) and rare outcomes (rare diseases in sub populations) and understanding why certain treatments (chemotherapy) benefit a small sub population and harm another sub population. To better illustrate this point we introduce the following mixture model for the population (congregation of big data).

$$f(X\,;\,\theta) = w_1 f_1(x,\theta_1) + w_2 f_2(x,\theta_2) + \cdots + w_k f_k(x,\theta_k) \qquad \cdots (1.1)$$

where $w_i \geq 0$, represents weights (proportion) of the i[th] subgroup and $f_i(x,\theta_i)$ is the probability of the response of the i^{th} subpopulation given in the covariates X with θ_i as a parameter vector. In practice, many subpopulations are rarely observed (w_i is very small). When the sample size n is moderate, nw_i can be small, making it infeasible to infer the covariate-dependent parameters due to lack of information. However, because big data characterized by large sample size n, the sample size nw_i for the i[th] population can be moderately large even if w_i is very small. This enables us to more accurately infer about the parameters of the subpopulation. In short, the main advantage brought by big data is to understand heterogeneity of subpopulation such as the benefits of certain personalized treatments, which are infeasible when sample size is small or moderate.

Big data also allows us to unveil weak commonality across whole population, thanks to large sample sizes. For example for benefit of one peg of red wine per night on heart can be difficult to access without large samples. Similarly, health risks to exposure of certain environmental factors can only be more convincingly evaluated when the sample sizes are sufficiently large.

Besides the afore mentioned advantages the heterogeneity of big data also possess significant challenges to statistical inference. Inferring the mixture model (1.1) for large data set requires sophisticated statistical and computational methods. In low dimensions, standard technique like Expectation Maximization (E-M) algorithm for finite mixture models can be applied. In high dimensions, we need to carefully regularize the estimating procedure to avoid over fitting or noise accumulation and to device good computational algorithm which is a challenging and open research area for statistical computational scientists. The maximum number of attributes considered by us in this work is nine. Hence comes under low dimensionality and therefore we have adopted E-M algorithm to estimate parameters and details will follows.

II. METHODOLOGY OF MODELLING AND ESTIMATION

Srinivas et al (2014) have applied five clustering methods on six datasets and they been validated and the resultant optimal clusters based on seven validation techniques. This section is devoted for fitting Gaussian Mixture Model (GMM) (assuming the weights/ proportionality parameter are unknown) for the optimal clusters. The methodological details using the following notations are as under

K	:	Number of clusters
$\boldsymbol{N_p(\mu_i, \Sigma_i)}$	:	Multivariate Normal population of ith cluster with mean vector μ_i and variance covariance Σ_i
$\boldsymbol{w_i}$	:	Weight / Proportionality parameter of ith cluster.

Using the above notations, Let GMM be

$$f(X, \theta) = \sum w_i f_i(X, \theta_i) \quad \cdots \quad (2.1)$$

where $\theta_i : \{\mu_i, \Sigma_i\}$; $w_i \geq 0, \sum w_i = 1$

The above model is fitted to the resultant clusters based on Expectation and Maximization (E-M). E-M algorithm is basically method of maximum likelihood in estimating the parameters of GMM. Hence can be viewed as likelihood maximization via the E-M algorithm.

There are a couple of issues related to ML estimation. One practical problem related to the estimation of parameters in finite mixture model is troublesome optimization. The form of likelihood function in case of GMM is a bit complicated and in most of the situations a multimodal, making it analytically intractable to obtain ML estimators in a closed form especially complicated multi parameter situation as of GMM and hence it is the primary tool in finite mixture models and model based clustering.

The E-M algorithm is an iterative process consisting of two steps first step consisting of finding expectation and second step is to maximise the expectation. At Eth step of the sth iteration, the posterior probabilities are calculated as

$$w_{ik}^{(s)} = Prob\{X_i \in k^{th}\ cluster | X_i, \theta^{(s-1)}\} = \frac{w_k^{(s-1)} f_k\left(x_i; \theta_k^{(s-1)}\right)}{\sum_{j=1}^{k} w_j^{(s-1)} f_k\left(x_i; \theta_k^{(s-1)}\right)} \cdots (2.2)$$

while the Mth step maximizes the expected conditional complete likelihood. Historically denoted as Q-function, with respect to vector parameter $\theta : Q(\theta;\ \theta^{(s-1)}, x_1, x_{2\ldots} x_n)$ where Q is given by

$$Q(\theta;\ x_1, x_2, \ldots, x_n) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\left[log|\Sigma_i| + (x_i - \mu_j)\Sigma_j^{-1}(x_i - \mu_j)\right] + \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\ log w_j - \frac{pn}{2} log 2\pi \cdots (2.3)$$

The iteration E and M steps until the relative increase in the likelihood function is no longer appreciable. In this context we have adopted Aitken's stopping rule,

$$\left|l^{(s+1)} - l^{(s)}\right| < \in \cdots (2.4)$$

where $\in$ is the tolerance level and $l^{(s)}$ is the Aitken's accelerated estimate of the limiting value such that

$$l^{(s+1)} = l^{(s)} + \frac{l^{(s+1)} - l^{(s)}}{1 - \frac{l^{(s+1)} - l^{(s)}}{l^{(s)} - l^{(s-1)}}} \cdots (2.5)$$

The expectation for updated vector parameter $\theta^{(s)}$ at the M- step may not be necessarily be of the closed form in which case Q function should be maximised numerically. Multivariate Normal mixtures $[\boldsymbol{N_p(\mu,\ \Sigma)}]$ are used in this case whose density function of the j^{th} component is given by

$$\phi(x{:}\,\mu_j, \Sigma_j) = 2\pi^{-\frac{p}{2}}|\Sigma_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \mu_j)'\Sigma_k^{-1}(x - \mu_j)\right\} \cdots (2.6)$$

The E- step consists of updating posterirori probabilities given the current parameter estimates of the vector parameter $\theta^{(s-1)}$ is given by

$$p_{ij}^{(s)} = \frac{w_j^{(s-1)}\phi(x_i, \mu_j^{(s-1)}, \Sigma_j^{(s-1)})}{\sum_{j=1}^{k} w_j^{(s-1)}\phi(x_i, \mu_j^{(s-1)}, \Sigma_j^{(s-1)})} \cdots (2.7)$$

In this case the M-step provides the convenient closed form solutions

$$w_j^{(s)} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(s)} \; ; \; \mu_j^{(s)} = \frac{\sum_{i=1}^{n} p_{ij}^{(s)} x_i}{\sum_{i=1}^{n} p_{ij}^{(s)}} \quad \cdots (2.8)$$

and

$$\Sigma_j^{(s)} = \frac{\sum_{i=1}^{n} p_{ij}^{(s)}\left(x_i - \mu_j^{(s)}\right)\left(x_i - \mu_j^{(s)}\right)'}{\sum_{i=1}^{n} p_{ij}^{(s)}} \quad \cdots (2.9)$$

III. RESULTS OF E-M

The E-M algorithm explained in section (2) is implemented using Matlab R2011b and the results for various datasets are given in Tables T3.1 through T3.6

The fitted GMM for Iris data is given by

$$f(X\,;\,\theta) = w_1\, N_4(\mu_1, \Sigma_1) + w_2\, N_4(\mu_2, \Sigma_2) + w_3\, N_4(\mu_3, \Sigma_3) \quad \cdots (3.1)$$

where the values of parameters in (3.1) are given in Table T3.1

Table T3.1: Estimation of Parameters of Iris data set

Class	**Weight**	$\boldsymbol{\mu}$	**Σ (Symmetricmatrix)**
1	w_1=0.333333	$\mu_1 = \begin{pmatrix} 1.17366 \\ 0.41669 \\ -0.50426 \\ -1.08608 \end{pmatrix}$	$\Sigma_1 = \begin{matrix} 0.0024 & & & \\ -0.0041 & 0.0075 & & \\ 0.0011 & -0.0031 & 0.0037 & \\ 0.0006 & -0.0003 & -0.0017 & 0.0014 \end{matrix}$
2	w_2=0.344615	$\mu_2 = \begin{pmatrix} 1.18837 \\ -0.39935 \\ 0.34574 \\ -1.13476 \end{pmatrix}$	$\Sigma_2 = \begin{matrix} 0.0030 & & & \\ 0.0004 & 0.0062 & & \\ -0.0049 & -0.0038 & 0.0097 & \\ 0.0015 & -0.0028 & -0.0010 & 0.0023 \end{matrix}$
3	w_3=0.322052	$\mu_3 = \begin{pmatrix} 1.06948 \\ -0.61705 \\ 0.59730 \\ -1.04973 \end{pmatrix}$	$\Sigma_3 = \begin{matrix} 0.0034 & & & \\ -0.0001 & 0.0046 & & \\ -0.0044 & -0.0013 & 0.0060 & \\ 0.0009 & -0.0034 & -0.0003 & 0.0028 \end{matrix}$

The fitted GMM for **Glass** data is given by

$$f(X\,;\,\theta) = w_1\, N_9(\mu_1\,,\Sigma_1) + w_2\, N_9(\mu_2\,,\Sigma_2) + w_3\, N_9(\mu_3\,,\Sigma_3) + w_4\, N_9(\mu_4\,,\Sigma_4) + w_5\, N_9(\mu_5\,,\Sigma_5) + w_6\, N_9(\mu_6\,,\Sigma_6) \quad \cdots (3.2)$$

where the values of parameters in (3.2) are given in Table T3.2

Table T3.2: Estimation of Parameters of Glass data set

Class	Weight	μ	Σ
1	w_1=0.065248	$\mu_1 = \begin{pmatrix} -0.41084 \\ 0.08380 \\ -0.44311 \\ -0.41265 \\ 2.60728 \\ -0.45387 \\ -0.02467 \\ -0.47419 \\ -0.47176 \end{pmatrix}$	$\Sigma_1 = \begin{array}{ccccccccc} 0.0000 \\ 00.000 & 0.0023 \\ -0.0001 & -0.0008 & 0.0008 \\ 0.0000 & -0.0002 & 0.0001 & 0.0003 \\ 0.0000 & -0.0004 & 0.0001 & 0.0001 & 0.0001 \\ 0.0001 & -0.0008 & -0.0001 & 0.0001 & 0.0002 & 0.0001 \\ 0.0000 & 0.0001 & 0.0000 & -0.0002 & -0.0001 & -0.0006 & 0.0011 \\ 0.0000 & -0.0001 & -0.0001 & 0.0000 & 0.0000 & 0.0001 & 0.0000 & 0.0000 \\ 0.0000 & -0.0001 & -0.0001 & 0.0000 & 0.0000 & 0.0001 & -0.0001 & 0.0000 & 0.0000 \end{array}$
2	w_2=0.164466	$\mu_2 = \begin{pmatrix} -0.41898 \\ 0.10917 \\ -0.33820 \\ -0.44156 \\ 2.60780 \\ -0.47815 \\ -0.07453 \\ -0.48382 \\ -0.48172 \end{pmatrix}$	***1.00e-03 **** $\Sigma_2 = \begin{array}{ccccccccc} 0.0134 \\ -0.0284 & 0.3879 \\ 0.0034 & -0.1249 & 0.7192 \\ 0.0181 & -0.0113 & -0.3380 & 0.3194 \\ 0.0121 & -0.0653 & 0.0271 & 0.0283 & 0.0221 \\ 0.0046 & -0.0276 & -0.0730 & 0.0621 & 0.0138 & 0.0423 \\ -0.0526 & -0.0290 & -0.2033 & -0.1319 & -0.0749 & -0.0488 & 0.6820 \\ 0.0154 & -0.0390 & -0.0134 & 0.0305 & 0.0167 & 0.0123 & -0.0678 & 0.0248 \\ 0.0149 & -0.0615 & 0.0039 & 0.0238 & 0.0211 & 0.0152 & -0.0729 & 0.0215 & 0.0349 \end{array}$
3	w_3=0.559976	$\mu_3 = \begin{pmatrix} -0.41467 \\ 0.07790 \\ -0.33027 \\ -0.42064 \\ 2.61871 \\ -0.45431 \\ -0.12114 \\ -0.47902 \\ -0.47655 \end{pmatrix}$	***1.00e-03**** $\Sigma_3 = \begin{array}{ccccccccc} 0.0064 \\ -0.0226 & 0.2271 \\ -0.0051 & 0.0269 & 0.0930 \\ 0.0024 & -0.0231 & -0.0102 & 0.0729 \\ 0.0042 & -0.0325 & 0.0013 & 0.0101 & 0.0074 \\ 0.0057 & -0.0386 & -0.0049 & 0.0110 & 0.0076 & 0.0156 \\ -0.0007 & -0.0708 & -0.0836 & -0.0607 & -0.0085 & -0.0076 & 0.2254 \\ 0.0061 & -0.0262 & -0.0081 & 0.0022 & 0.0048 & 0.0065 & 0.0008 & 0.0089 \\ 0.0047 & -0.0391 & -0.0082 & -0.0036 & 0.0065 & 0.0057 & 0.0067 & 0.0061 & 0.0222 \end{array}$

1.003e-03* indicates that each of (i, j)th element of Σ_j is to be multiplied with 1.003e-03

Table T3.2: (Contd.)

4	w_4=0.051248	$\mu_4 = \begin{pmatrix} -0.41586 \\ 0.04573 \\ -0.48097 \\ -0.42298 \\ 2.59465 \\ -0.46828 \\ -0.09306 \\ -0.46819 \\ -0.47716 \end{pmatrix}$	$\Sigma_4 = \begin{matrix} 0.0001 \\ 0.0001 & 0.0019 \\ 0.0001 & 0.0001 & 0.0001 \\ 0.0000 & 0.0001 & 0.0000 & 0.0003 \\ 0.0000 & -0.0001 & 0.0000 & 0.0002 & 0.0001 \\ 0.0000 & 0.0000 & 0.0001 & 0.0002 & 0.0001 & 0.0001 \\ -0.0001 & -0.0012 & -0.0002 & -0.001 & -0.0005 & -0.0005 & 0.0038 \\ -0.0002 & -0.0008 & -0.0002 & 0.0002 & 0.0002 & 0.0000 & -0.0002 & 0.0012 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0001 & -0.0001 & 0.0000 \end{matrix}$
5	w_5=0.131025	$\mu_3 = \begin{pmatrix} -0.41264 \\ 0.13889 \\ -0.46559 \\ -0.38485 \\ 2.60787 \\ -0.45375 \\ -0.11920 \\ -0.43452 \\ -0.47621 \end{pmatrix}$	$\Sigma_5 = \begin{matrix} 0.0001 \\ 0.0001 & 0.0010 \\ 0.0000 & 0.0001 & 0.0010 \\ 0.0000 & -0.0004 & -0.0005 & 0.0005 \\ 0.0000 & -0.0002 & 0.0000 & 0.0000 & 0.0000 \\ -0.0003 & -0.0011 & -0.0006 & 0.0004 & 0.0002 & 0.0042 \\ 0.0001 & 0.0000 & 0.0001 & 0.0002 & 0.0000 & -0.0015 & 0.0011 \\ 0.0001 & 0.0003 & -0.0001 & -0.0002 & 0.0000 & -0.0001 & 0.0000 & 0.0009 \\ 0.0001 & 0.0001 & 0.0000 & 0.0000 & 0.0000 & -0.0003 & 0.0001 & 0.0001 & 0.0001 \end{matrix}$
6	w_6=0.028037	$\mu_3 = \begin{pmatrix} -0.41959 \\ 0.12869 \\ -0.35796 \\ -0.39714 \\ 2.6738 \\ -0.44012 \\ -0.20467 \\ -0.44807 \\ -0.48393 \end{pmatrix}$	$\Sigma_6 = \begin{matrix} 0.0001 \\ 0.0000 & 0.0002 \\ 0.0000 & 0.0001 & 0.0005 \\ -0.0002 & -0.0001 & -0.0003 & 0.0004 \\ 0.0000 & -0.0001 & 0.0000 & 0.0000 & 0.0000 \\ -0.0001 & -0.0003 & 0.0000 & 0.0002 & 0.0001 & 0.0006 \\ 0.0003 & 0.0004 & -0.0001 & -0.0004 & -0.0002 & -0.0008 & 0.0016 \\ -0.0002 & -0.0003 & -0.0002 & 0.0004 & 0.0001 & 0.0005 & -0.0011 & 0.0010 \\ 0.0001 & 0.0000 & 0.0000 & -0.0002 & 0.0000 & -0.0001 & 0.0003 & -0.0003 & 0.0001 \end{matrix}$

The fitted GMM for **Breast Cancer** data is given by $f(X\,;\,\theta) = w_1\, N_9(\mu_1\,,\Sigma_1) + w_2\, N_9(\mu_2\,,\Sigma_2) \quad \cdots$ (3.3)

where the values of parameters in (3.3) are given in Table T3.3

Table T3.3: Estimation of Parameters of Breast Cancer data set

Class	**Weight**	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$
1	w_1=0.467829	$\mu_1 = \begin{pmatrix} 6.198957 \\ 5.403595 \\ 5.373185 \\ 4.626818 \\ 4.675772 \\ 6.314781 \\ 5.097618 \\ 4.926455 \\ 2.259891 \end{pmatrix}$	$\Sigma_1 = \begin{array}{lllllllll} 8.0469 & & & & & & & & \\ 4.1208 & 9.9836 & & & & & & & \\ 4.2491 & 8.2096 & 9.4563 & & & & & & \\ 2.0704 & 5.3308 & 4.9610 & 10.8087 & & & & & \\ 1.9662 & 4.5866 & 4.1331 & 3.0780 & 6.2677 & & & & \\ 3.3767 & 4.1391 & 4.7442 & 5.0505 & 2.2339 & 12.9686 & & & \\ 2.5927 & 4.9241 & 4.6421 & 4.2724 & 2.5767 & 3.88770 & 6.6509 & & \\ 2.8977 & 5.5831 & 5.5558 & 4.1686 & 3.4107 & 2.62320 & 4.1332 & 11.8724 & \\ 1.3640 & 2.2651 & 2.0466 & 2.0726 & 2.0445 & 0.97640 & 0.9809 & 2.19350 & 5.4338 \end{array}$
2	w_2=0.532171	$\mu_2 = \begin{pmatrix} 2.851879 \\ 1.139708 \\ 1.303542 \\ 1.206956 \\ 1.932764 \\ 1 \\ 1.978602 \\ 1.056454 \\ 1 \end{pmatrix}$	$\Sigma_2 = \begin{array}{lllllllll} 2.5618 & & & & & & & & \\ 0.0983 & 0.1900 & & & & & & & \\ 0.1706 & 0.0969 & 0.5117 & & & & & & \\ 0.1622 & 0.0276 & 0.0045 & 0.3253 & & & & & \\ 0.0812 & 0.0280 & 0.0308 & 0.0192 & 0.1703 & & & & \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & & & \\ 0.0311 & 0.0005 & -0.0120 & -0.0358 & -0.0258 & 0.0000 & 0.7584 & & \\ 0.0378 & 0.0110 & -0.0009 & 0.0152 & 0.0145 & 0.0000 & 0.0229 & 0.0533 & \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{array}$

The fitted GMM for **Half moon** data is given by

$$f(X\,;\,\theta) = w_1\,N_2(\mu_1\,,\Sigma_1) + w_2\,N_2(\mu_2\,,\Sigma_2) \quad \cdots (3.4)$$

where the values of parameters in (3.4) are given in Table T3.4

Table T3.4: Estimation of Parameters of Half moon data set

Class	Weight	μ	Σ
1	w_1=0.306191	$\mu_1 = \begin{pmatrix}36.35161\\10.10747\end{pmatrix}$	$\Sigma_1 =$ 9.4738 10.9185 15.9469
2	w_2=0.693809	$\mu_2 = \begin{pmatrix}19.02563\\13.04561\end{pmatrix}$	$\Sigma_2 =$ 43.4646 -30.5666 53.0367

The fitted GMM for **Path based** data is given by

$$f(X\,;\,\theta) = w_1\,N_2(\mu_1\,,\Sigma_1) + w_2\,N_2(\mu_2\,,\Sigma_2) + w_3\,N_2(\mu_3\,,\Sigma_3) \cdots (3.5)$$

where the values of parameters in (3.5) are given in Table T3.5

Table T3.5: Estimation of Parameters of Path based data set

Class	Weight	μ	Σ
1	w_1=0.068621	$\mu_1 = \begin{pmatrix}14.98454\\29.80523\end{pmatrix}$	$\Sigma_1 =$ 15.2067 5.2301 2.2744
2	w_2=0.484207	$\mu_2 = \begin{pmatrix}26.55217\\16.72708\end{pmatrix}$	$\Sigma_2 =$ 10.5811 -1.0064 30.4583
3	w_3=0.447172	$\mu_3 = \begin{pmatrix}11.09528\\15.75203\end{pmatrix}$	$\Sigma_3 =$ 11.362 -0.34 16.3802

The fitted GMM for **Spiral** data is given by

$$f(X\,;\,\theta) = w_1\,N_2(\mu_1\,,\Sigma_1) + w_2\,N_2(\mu_2\,,\Sigma_2) + w_3\,N_2(\mu_3\,,\Sigma_3) \cdots (3.6)$$

where the values of parameters in (3.6) are given in Table T3.6

Table T3.6: Estimation of Parameters of Spiral data set

Class	Weight	μ	Σ
1	w_1=0.3453636	$\mu_1 = \begin{pmatrix}15.6274\\10.30082\end{pmatrix}$	$\Sigma_1 =$ 35.2652 -9.8373 16.9311
2	w_2=0.351857	$\mu_2 = \begin{pmatrix}25.30147\\17.48634\end{pmatrix}$	$\Sigma_2 =$ 14.2423 -2.0796 36.043
3	w_3=0.302781	$\mu_3 = \begin{pmatrix}13.56943\\21.91192\end{pmatrix}$	$\Sigma_3 =$ 31.2057 12.4647 19.8818

IV. VALIDATING THE ESTIMATED GMMs

The GMMs obtained in Section (3) is validated based on the following eight measures for validating clusters with respect to original class labels obtained by applying AMOC suggested by Srinivas et al (2014), they are (i) Accuracy (ACC), (ii) Precession (PRE), (iii) Sensitivity (SEN), (iv) Specificity (SPE), (v) Negative Predicted Value (NPV), (vi) F-Measure (FM), (vii) Mathew's Correlation Coefficient (MCC) and (viii) Overall Accuracy (OVA) and are presented in Table T4.1

Table T4.1: Estimated Accuracy measures of various data sets when GMM is fitted.

GMM	CLASS	ACC	PRE	SEN	SPE	NPV	FM	MCC	OVA (%)
IRIS	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	98.67
	2	0.9867	0.9615	1.0000	0.9800	1.0000	0.9804	0.9707	
	3	0.9867	1.0000	0.9600	1.0000	0.9804	0.9796	0.9701	
GLASS	1	0.9813	1.0000	0.7778	1.0000	0.9800	0.875	0.8731	95.79
	2	0.9813	0.8857	1.0000	0.9781	1.0000	0.9394	0.9308	
	3	0.9766	1.0000	0.9600	1.0000	0.9468	0.9796	0.9534	
	4	0.9813	0.6364	1.0000	0.9807	1.0000	0.7778	0.7900	
	5	0.9953	0.9643	1.0000	0.9947	1.0000	0.9818	0.9794	
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
BC	1	0.8999	0.7859	1.0000	0.8416	1.0000	0.8801	0.8133	89.99
	2	0.8999	1.0000	0.8416	1.0000	0.7859	0.914	0.8133	
HM	1	0.6595	1.0000	0.4816	1.0000	0.5020	0.6501	0.4917	65.95
	2	0.6595	0.5020	1.0000	0.4816	1.0000	0.6684	0.4917	
PB	1	0.9633	1.0000	0.6563	1.0000	0.9606	0.7925	0.7940	96.33
	2	0.9633	0.9247	1.0000	0.9333	1.0000	0.9609	0.9290	
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
SPIRAL	1	0.5353	0.3178	0.3208	0.6456	0.6488	0.3192	-0.0335	32.69
	2	0.5513	0.3211	0.3465	0.6493	0.6749	0.3333	-0.0041	
	3	0.5673	0.3438	0.3143	0.6957	0.6667	0.3284	0.0102	

COMMENTS

From Table T4.2 when GMMs are fitted for various data sets the following observations are noted.

- ✓ In case of Iris data GMM shows the highest overall accuracy of 98.67 % when compared to other data sets.
- ✓ Glass and Path based data exhibits on average 96% of accuracy.
- ✓ For Breast Cancer data the overall accuracy is 90% when compared to other data sets.
- ✓ Half moon data gave an accuracy of only 66%.
- ✓ The overall accuracy for the Spiral data is 32.69 % which is the least and this as expected because the data is not linearly separable.

REFERENCES

[1] Anil K. Jain (1989) Fundamentals of Digital Image Processing, PHI.

[2] Anderson, T.W. (1962). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

[3] Andrews, H.C. (1972). Introduction to Mathematical Techniques in Pattern Recognition, John Wiley & Sons, New York.

[4] Alex Berson, Stephen Smith, Kurt Thearling (2000) Building Data Mining Applications for CRM, Tata McGraw Hill.

[5] Arun. K Pujari (2001) Data Mining Techniques, University Press (India) Ltd.

[6] Broomhead D. S, Lowe David (1988) Multivariate functional interpolation and adaptive networks, Complex Systems 2: 321-355.

[7] Bandyopadhyay. S et al (1995). Pattern classification with genetic algorithms. Patterns Recognition Letters, Vol. 16(8) pp. 801-808.

[8] Bart Kosko (1997) Neural Networks and Fuzzy Systems: A Dynamical Systems approach to machine Intelligence, PHI.

[9] Bose. N. K, Liang. P (1998) Neural Network Fundamentals with Graphs, Algorithms and Applications, Tata McGraw Hill.

[10] Bhuvan K. C (2005) Multivariate Analysis and its Applications, New Central Book Agency (P) Ltd.

[11] Christopher M. Bishop (1995) Neural Networks for Pattern Recognition, Oxford University Press.

[12] David Hand, Heikki Mannila, Padhraic Smyth (2004) Principles of Data Mining, PHI.

[13] Ethem Alpaydin (2010) Introduction to machine Learning, MIT Press.

[14] Fisher R. A (1936) The use of Multiple measurements in Taxonomic Problems, Annals of Eugencies

[15] Fuknaga Keinosuke (1972) Introduction to Statistical Pattern Recognition, Academic Press.
[16] German B (1987) Glass identification Database, UCI Machine Learning Repository.
[17] Han Jiawei, Kamber Micheline (2001) Data Mining: Concepts and Techniques. Morgan Kaufmann
[18] James A. Anderson (1999) An Introduction to Neural Networks, PHI.
[19] Johnson Richard A., and Wichern Dean. W(2001). Applied Multivariate Statistical Analysis, Prentice Hall International Inc., Englewood Cliffs, New Jersey.
[20] Jang-J-S.R, Sun.T-C, Mizutani. E (2003) Neuro-Fuzzy and Soft Computing: A computational approach to learning and Machine Intelligence.
[21] James A Freeman, David M. Skapura (2013) Neural Networks Algorithms, Applications and Programming Techniques.
[22] Karteeka Pavan K, Allam Appa Rao, A V Dattatreya Rao, GR Sridhar (2008) Automatic generation of Merge factor for clustering Microarray Data, International journal of Computer Science and Network and Security,South Korea, vol.8,issue.9,pp127-131.
[23] Mc Lachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern. John Wiley, New York
[24] Mohamad H. Hassoun (1999) Fundamentals of Artificial Neural Networks, PHI.Margaret H. Dunham (2003) Data Mining, Introductory and Advanced Topics, Pearson Edition.
[25] Patrick, E. A. (1972) Fundamentals of Pattern Recognition, Prentice Hall, Englewood Cliffs, New Jersey.
[26] Prabhu C. S. R (2002) Data Warehousing, Concepts, Techniques, products and Applications, PHI.
[27] Richard, M. D and Lippmann, R. P (1991) Neural Network Classifiers Estimate Bayesian a posteriori Probabilities, Neural Computation 3, pp. 461-483
[28] Ripley B. D (1996) Pattern Recognition and Neural Networks, University of Oxford, Cambridge University Press.
[29] Robert S (2005) Pattern Recognition, Statistical Structural and Neural Approaches.
[30] Rajasekaran. S, Vijayalakshmi Pai. G. V (2006) Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Applications, PHI.
[31] Srinivas N, Dattatreya Rao. A. V, Karteeka Pavan. K (2014) Relative Performance of K-Means, Single Linkage and Affinity Propagation in Cluster Analysis TIJ's Research Journal of Science & IT Management (RJSITM).
[32] Volodymyr Melnykov and Ranjan Maitra (2010): Finite Mixture models and models-based clustering. Statistical surveys, Vol. 4 (2010), Pg. 80-116
[33] William H. Wolberg and O.L. Mangasarian: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196