IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 9, November 2014.
www.ijiset.com
ISSN 2348 – 7968

Relative Performance of Automatic Merging of Clusters (AMOC) and Self Organizing Map (SOM) in Cluster Analysis

N.Srinivas[1], A.V.Dattatreya Rao[2] and K. Karteeka pavan[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

Karteeka et al (2008) suggested AMOC and an attempt is made to study the relative performance of Automatic Merging of Clusters (AMOC) and Self Organizing Map (SOM) in clustering six public data sets viz., Iris, Glass, Breast Cancer, Half-moon, Path based and Spiral. The performance of clustering methods is studied based on seven validation techniques viz., Rand, Adjusted Rand, Error Rate, Silhouette, Davies-Bouldin, Dunn and CS Indices. The results obtained empirically and conclusions are summarized in section 6.

1. Introduction

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within same cluster and are dissimilar to the objects in other clusters. It is astounding to note the number of possible ways of arranging 25 observations into 5 groups is 2,436,684,974,110,751 and it is well known from the theory of combinatorics that the number of ways of arranging **n** observations into **m** groups is

$$\frac{1}{m!}\sum_{k=0}^{m}(-1)^{m}\binom{m}{k}k^{n}$$

This is the motivating factor to develop new clustering algorithms or to study the relative performance of the existing clustering procedures for the massive Exabyte or Petabytes of data to be processed with the popularity of various social networks or scientific data such as genome, remote sensing etc. The present data scientists are confronted to resolve the issues, as a result of big data or high dimensional data or big data coupled with high dimensional data. In this paper a modest attempt is made to study relative performance of two different clustering algorithms based on six data sets with respect to seven validity measures.

This paper is divided into six sections. Section 2 is devoted to describe various datasets used in the present data. Section 3 deals with brief description of the algorithms. Validation measures are presented in section 4. Section 5 and 6 are respectively devoted for presenting methodology of the study and observations followed by the conclusions.

2. Description of data sets used

Iris Data [Fisher R.A. (1936)] described three varieties of Iris flower (Setosa, Versicolor, and Virginica) each having four features (Sepal length, Sepal width, Petal length and Petal width) for the purpose of Pattern recognition.

Glass Data [German B (1987)] has six classes (Building window float, Building window non-float, Vehicle windows float, Containers, Table ware and Headlamps) each having nine attributes (Refractive index, Sodium, Magnesium, Aluminium, Silicon, Potassium, Calcium, Barium and Iron)

Breast Cancer Data [Wolberg (1991)] The breast cancer data was obtained from the university of Wisconsin Hospital. The data set contains two classes Benign and Malignant. having nine features - Clump Thickness, Uniformity of Cell size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nucleoli, Bland Chromatin, Normal Nucleoli and Mitoses.

Synthetic Datasets: These synthetic data sets are the simulated data sets of multivariate normal $N_p(\mu, \Sigma)$ for specified values of μ and Σ. Half moon, Path based and Spiral data sets are specific Bivariate Normal samples.

Parameters of **Half moon** are

$$\mu_1 = \begin{pmatrix} 27.7150 \\ 9.0254 \end{pmatrix}; \mu_2 = \begin{pmatrix} 14.7010 \\ 21.0253 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 69.0757 & 2.5356 \\ & 16.9593 \end{pmatrix}; \sum_2 = \begin{pmatrix} 51.7205 & -1.9333 \\ & 12.8283 \end{pmatrix}$$

those of **Path based** are

$$\mu_1 = \begin{pmatrix} 19.315 \\ 19.0632 \end{pmatrix}; \mu_2 = \begin{pmatrix} 12.5119 \\ 15.9995 \end{pmatrix}; \ \mu_3 = \begin{pmatrix} 24.8941 \\ 16.2113 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 110.5167 & -3.1809 \\ & 76.1151 \end{pmatrix}; \sum_2 = \begin{pmatrix} 5.1379 & -0.1683 \\ & 6.5701 \end{pmatrix}; \sum_3 = \begin{pmatrix} 5.5303 & -0.2709 \\ & 7.7563 \end{pmatrix}$$

and the parameters of **Spiral** dataset are

$$\mu_1 = \begin{pmatrix} 20.4252 \\ 16.3257 \end{pmatrix}; \mu_2 = \begin{pmatrix} 17.3614 \\ 17.9324 \end{pmatrix} \ \mu_3 = \begin{pmatrix} 17.5231 \\ 14.7901 \end{pmatrix}$$

$$\sum_1 = \begin{pmatrix} 50.2159 & 9.8161 \\ & 49.0839 \end{pmatrix}; \ \sum_2 = \begin{pmatrix} 60.8102 & -2.3806 \\ & 37.3353 \end{pmatrix} \sum_3 = \begin{pmatrix} 43.9809 & -6.4972 \\ & 50.9929 \end{pmatrix}$$

Note: Datasets are downloaded from the UCI Machine Learning Repository.

3. Clustering methods Studied

Here two clustering techniques are applied on the data sets described in the earlier section**.** They are Automatic Merging of Clusters (AMOC) and Self Organizing Map (SOM). Of the two, SOM is well known and hence the algorithm of AMOC only is briefly presented as under

1. Initialize k_{max}, number of clusters to the square root of total number of objects
2. Assign k_{max} objects randomly to the cluster centroids
3. Find the clusters using k-means
4. Compute Rand index
5. Find a cluster that has least probability and merge with its closest cluster. Recompute centroids, Rand index and decrement the number of clusters by one. If the newly computed Rand index is greater than the previous Rand index, then update and Index, number of clusters and cluster centroids with the newly computed values.
6. If step 5 has been executed for each and every cluster, then go to step7, otherwise go to step5.
7. If there is no change in number of clusters, then stop, otherwise go to step2

The AMOC, which automatically finds near optimal clusters. The algorithm requires the input parameter, the number of initial clusters which is a possible large number (any $k < m$, where k is the number of clusters and m is the number of elements) and the algorithm is not sensitive to this input parameter. The technique is a two-phase iterative procedure, which attempts to determine the appropriate clusters and improve the overall quality of the partitions. In the first phase, it produces clusters for a large k.

In the second phase, iteratively a low probability cluster is merged with its closest cluster using a validation technique.

The clustering algorithms are implemented in Matlab R2011b and the code is available with the first author.

4. Choosing the best Clustering Technique

We have used the following well known cluster validation indices to select the best clustering algorithm in a given situation, (a) Rand Index (RI), (b) Adjusted Rand Index (ARI), (c) Error Rate(ER), (d) Silhouette Index (SI), (e) Davies-Bouldin Index (DBI), (f) Dunn Index (DI) and (g) CS Index (CSI). Among the above indices RI, ARI ranges between 0 and 1 and we choose that method for which these values are nearer to 1. ER is expressed in percentages and we choose that method for which ER is minimum. SI for data point varies from -1 to 1, and hence we choose the method for which the average SI for all the points is maximum. We prefer DBI to be as minimum as possible as it denotes compactness. As DI is a measure of separation and varies between 0 to infinity, we prefer if it maximum. Finally CSI is a ratio of measures of within to between cluster distances and hence it should be as minimum as possible.

5. Methodology for the Study of Relative Performance

In this section we briefly describe the methodology of deciding relative performance of an algorithm when k, the number of clusters is known and unknown. In fact for the data sets chosen for this study, the number of classes are known before hand implying that k is specified. However to see the relative performance of the algorithm in deciding number of optimal clusters with respect to a specified validity measure is worked out presuming k is unknown.

Since the number of classes for each dataset are known, taking the **original dataset** (The size of the datasets respectively are: Iris150 x 4, Glass 214 x 9, Breast Cancer 699 x 9, Half moon 373 x2, Path based 300 x 2, Spiral 312 x 2) visual representation of the data with respect to label of the classes is plotted taking first two attributes as x axis and y axis respectively, and the graphs are presented in Fig. 1 through 6 in the appendix.

After implementing the two chosen algorithms for known k (k is taken as the true number of classes of the given dataset which is known a priori, k=3, 6, 2, 2, 3 and 3 respectively for Iris, Glass, Breast Cancer, Half moon, Path based and Spiral datasets) the clustering structures are plotted as mentioned in the earlier paragraph with respect to each algorithm for all six datasets and the graphs are presented in Fig. 7 through 12 for AMOC, Fig. 13 through 18 for SOM in the Appendix.

Since the number of classes are known as mentioned above, validity measures are tabulated with respect to the algorithms fixing a particular dataset and are presented in Table T1 through T6. The best algorithm with respect to each validity measure is shown in bold figures in the tables T1 through T6.

Presuming k (the number of clusters in the population of each dataset) is unknown optimum number of clusters is decided in comparison with those of the population with respect to a given validity measure for a given dataset and for each algorithm and these results are presented in Tables T7 through T9 and are presented in the appendix. Given a dataset and an algorithm, the best validity measure to be chosen is indicated in bold figures in tables T7 through T9

6. Observations and Conclusions

The observations from various tables and consolidated suggestive conclusions are presented in this section.

6.1 Observations

- From Table T7 the observations are as follows
  - ✓ For Iris data set when AMOC is used indices RI, ARI identified the correct number of 3 clusters.
  - ✓ For glass data both SI and DBI index selects the correct number of clusters.
  - ✓ For Breast Cancer data set all the 6 validity indices gives the optimal number of clusters as 2.
  - ✓ For Half moon data all validity indices selects exact number of clusters.
  - ✓ For Path Based dataset SI, DBI, DI and CSI selects correct number of clusters.
  - ✓ In case of Spiral no index selects accurate number of clusters
- From Table T8 the observations are listed as follows
  - ✓ For Iris data only SI selects 3 clusters which is the correct number of clusters.
  - ✓ For Glass CSI gaves the optimal number of clusters as 6.
  - ✓ For Breast Cancer data all indices show the correct number of clusters.
  - ✓ For Half moon data RI, SI, DI and CSI selects 2 clusters which is the optimal number of correct clusters
  - ✓ For Path Based data AR, SI and CSI validity index gives the correct number of clusters.
  - ✓ In case of Spiral data SI and DI index gives the correct number of clusters as 3 which are the optimal clusters
- The final observations from tables T1 through T6 are listed as under
  - ✓ For real data set Iris AMOC suits well
  - ✓ In case of Glass data set AMOC is best with minimum error rate.
  - ✓ For real dataset Breast Cancer the clustering algorithms both algorithms perform equally well.
  - ✓ For artificial dataset Half moon and Pathbased only AMOC gave the best performance in clustering solution
  - ✓ The performance of AMOC algorithm is better when compared to SOM.

6.2 Suggestive Conclusions

As the data scientists, ultimately we need to recommend the algorithm to be used for a given set and for a given a validity measure for the use of the practitioners as a thumb rule and this section serves this purpose.

Notations used in various tables to present the final conclusions are explained below

- ➢ A and S respectively indicate the algorithms Automatic Merging of Clusters (AMOC) algorithm and Self Organizing Map (SOM)
- ➢ YES indicates that all algorithms are equally good in terms of convergence of k, number of clusters obtained using a particular algorithm to true number of clusters in the population.
- ➢ NO indicates that none of the algorithms is converging.

✓ The ascending order performance of algorithms for a given data set and for a given validity measure assuming k is known (taking k= true number of clusters in the population) is tabulated in Table CT1 as follows.

Table CT1 *Ascending order of performance of algorithms when k is known*

	DATASET					
Validity Index	**Iris**	**Glass**	**Breast Cancer**	**Half moon**	**Path Based**	**Spiral**
RI	AS	SA	AS	AS	AS	AS
ARI	AS	AS	AS	AS	AS	AS
SI	AS	AS	AS	AS	AS	AS
DBI	AS	AS	SA	SA	SA	AS
DI	AS	AS	SA	SA	SA	SA
CSI	AS	AS	AS	AS	SA	AS
ER %	AS	AS	AS	AS	AS	SA

***Interpretation:* For example for Iris data when Rand Index is chosen the order of performance of alg1orithms are AMOC and SOM with respect to maximum value of Rand Index and other entries in the table can be interpreted according to the optimal value of the Indices when k is known.**

- ✓ The ascending order of performance of algorithms for a given data set and for a given validity measure assuming k is unknown (varying k from 2 to 10) is tabulated in Table CT2 as follows.

Table CT2 *Ascending order of performance of algorithms when k is unknown*

	DATASET					
Validity Index	**Iris**	**Glass**	**Breast Cancer**	**Half moon**	**Path Based**	**Spiral**
RI	AS	AS	YES	YES	NO	NO
ARI	AS	SA	YES	AS	SA	NO
SI	SA	SA	YES	YES	YES	SA
DBI	AS	AS	YES	AS	AS	AS
DI	NO	NO	YES	YES	AS	SA
CSI	NO	SA	YES	YES	YES	NO
ER %	AS	NO	YES	YES	NO	NO

***Interpretation:* For example for Iris data when Rand Index is chosen the order of performance of alg1orithms are AMOC and SOM with respect to maximum value of Rand Index and other entries in the table can be interpreted according to the optimal value of the Indices when k is unknown.**

- ✓ The ascending order of performance of various validity measures ((*denoting R, A, S, B, D, C and E respectively Rand, Adjusted Rand, Silhouette, Davies-Bouldin, Dunn, CS and Error Rate*) for a given dataset and the algorithm is tabulated in CT3 for k unknown as follows.

Table CT3 *Ascending order of performance of validity measures when k is unknown*

	DATASET					
Algorithm	**Iris**	**Glass**	**Breast Cancer**	**Half Moon**	**Path Based**	**Sprial**
AMOC	RAEDSC	SDRAEDC	YES	YES	SBDCARE	NO
SOM	SBDRAE	CABRED	YES	RSDCEA	ASCBDRE	SDCARBE

Interpretation: For example for IRIS when k-means algorithm is chosen the order of performance of validity measures are Rand Index, Adjusted Rand Index, Error Rate, Dunn, Davies-Bouldin Index, Silhouette, and CS measure and similarly other entries in the table can be interpreted.

REFERENCES

1. Anderson, T.W. (1962). An Introduction to Multivariate Statistical Analysis. Wiley, New York.
2. Andrews, H.C (1972). Introduction to Mathematical Techniques in Pattern Recognition, John Wiley & Sons, New York.
3. Arjun K Pujari (2001) Data Mining Techniques. University Press, India
4. Bandyopadhyay. S et al (1995). Pattern classification with genetic algorithms. Patterns Recognition Letters, Vol. 16(8) pp. 801-808.
5. Bose N. K and Liang P (1998) Neural Network Fundamentals with Graphs, Algorithms and Applications. Tata McGraw-Hill Publishing, New Delhi, India
6. Christopher M. Bishop (1995).Neural Networks for Pattern Recognition. Oxford university press.
7. Delbert Dueck (2009) Affinity Propagation: Clustering data by passing messages.
8. Fisher R. A (1936) The use of Multiple measurements in Taxonomic Problems, Annals of Eugencies
9. Fukunaga K. (1990). Introduction to Statistical Pattern Recognition. Academic Press, 2nd edition.
10. German B (1987) Glass identification Database, UCI Machine Learning Repository.
11. Jang, Sun and Mizutani (2003) Neuro-Fuzzy and Soft Computing. PHI publications, New Delhi, India
12. Johnson Richard A, and Wichern Dean W(2001). Applied Multivariate Statistical Analysis, Prentice Hall International Inc., Englewood Cliffs, New Jersey.
13. 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.
14. Mangasarian O. L and Wolberg W. H: "Cancer diagnosis via linear programming", SIAM a News, Volume 23, Number 5, September 1990, pp 1 & 18
15. Mclachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern John Wiley, New York
16. Patrick, E. A. (1972) Fundamentals of Pattern Recognition, Prentice Hall, Englewood Cliffs, New Jersey.
17. Precha Thavikulwat (2008) Affinity Propagation: A Clustering Algorithm for Computer- Assisted Business Simulations and Experiential Exercises, Developments in Business Simulation and Experiential Learning, Volume 35.
18. Shailendra Kumar S, J.L. Rana and Jain. R. C. Fast Affinity Propagation Clustering based on Machine Learning, IJCSI International Journal of Computer Science Issues, Vol. 10
19. 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).
20. 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.

APPENDIX

Visual Representation of Data Sets

1. Visual representation of the **Original data sets**

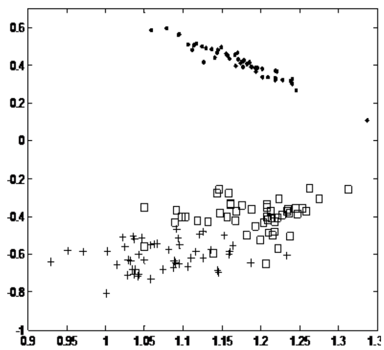

Fig 1 Plot of Iris data

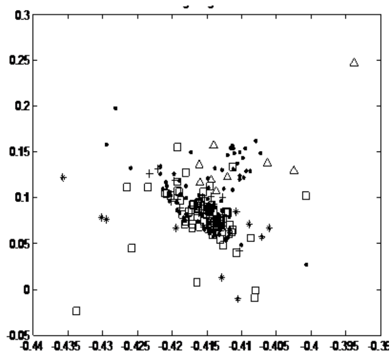

Fig 2 Plot of Glass data

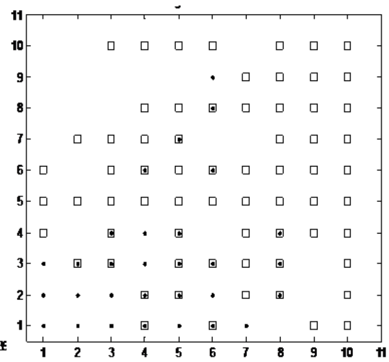

Fig 3 Plot of BC data

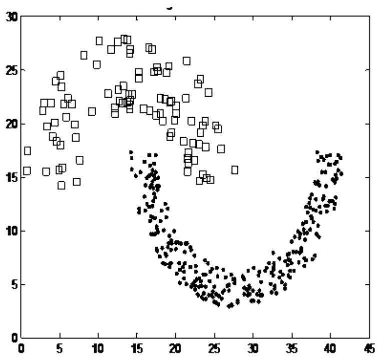

Fig 4 Plot of HM data

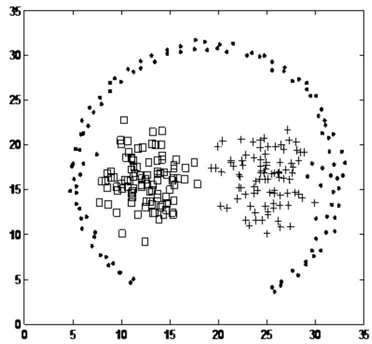

Fig 5 Plot of PB data

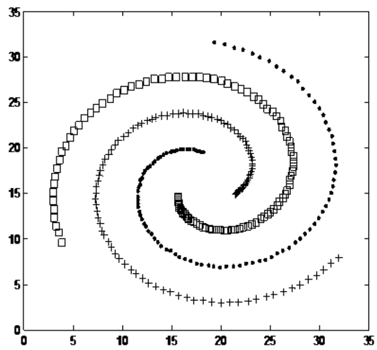

Fig 6 Plot of Spiral data

2. The clustering structures for known k with respect to **AMOC** algorithm.

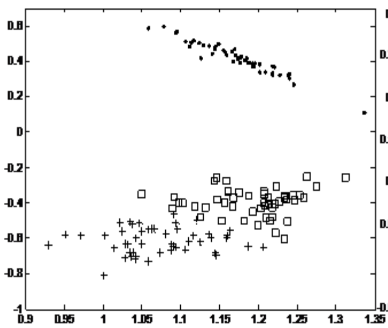

Fig 7 Plot of Iris data

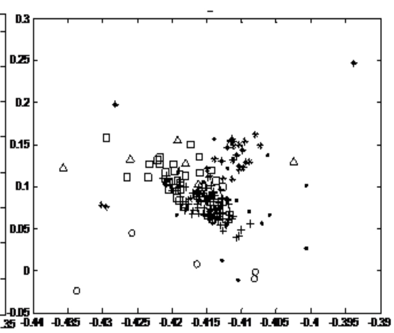

Fig 8 Plot of Glass data

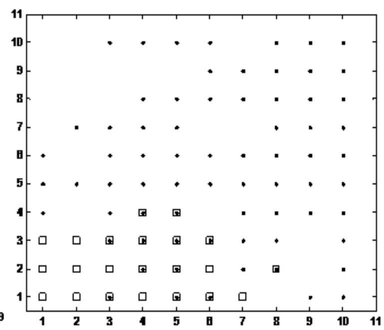

Fig 9 Plot of Breast Cancer

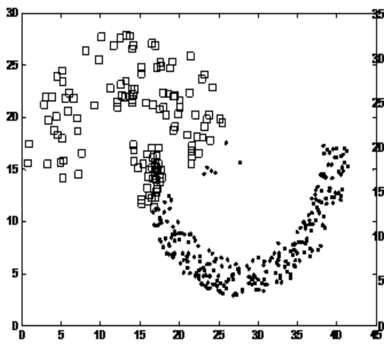

Fig 10 Plot of HM data

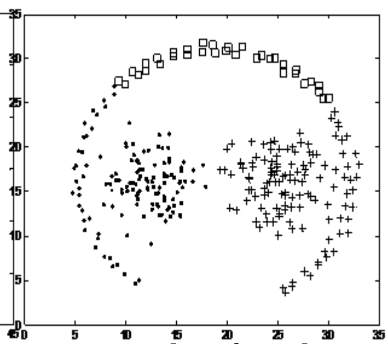

Fig 11 Plot of PB data

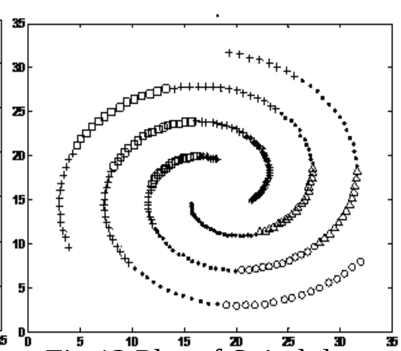

Fig 12 Plot of Spiral data

3. The clustering structures for known k with respect to **SOM** algorithm.

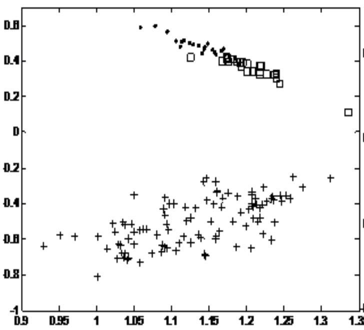

Fig 13 Plot of Iris data

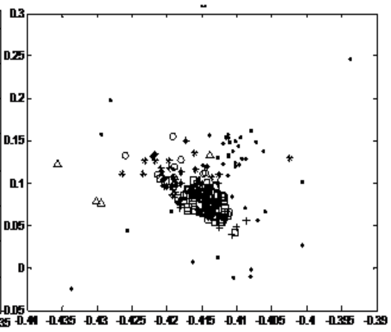

Fig 14 Plot of Glass data

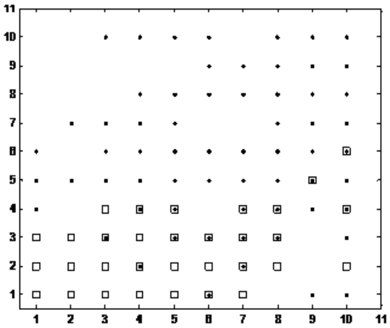

Fig 15 Plot of Breast Cancer

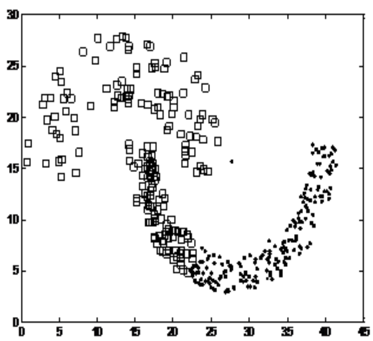

Fig 16 Plot of HM data

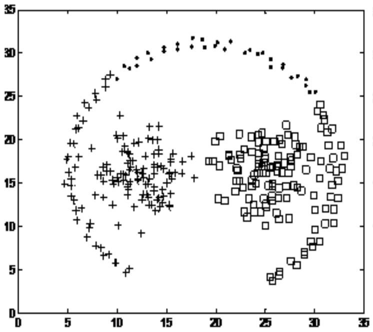

Fig 17 Plot of PB data

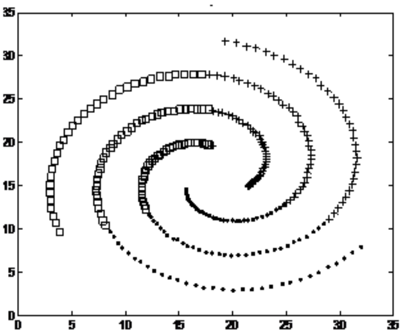

Fig 18 Plot of Spiral data

Tabulated Results obtained in this study

1. Validity measures with respect to the algorithms.

Table T1 *Validity measures for Iris data of various algorithms*

IRIS	**Validity Measure**						
Algorithm	**RI**	**ARI**	**SI**	**DBI**	**DI**	**CSI**	**ER (%)**
AMOC	**0.9495**	**0.8857**	**0.8055**	**0.4335**	**2.9127**	**0.6177**	**4.000**
SOM	0.7206	0.4394	0.6394	0.5826	0.7817	0.7524	50.000

Table T2 *Validity measures for Glass data of various algorithms*

GLASS	**Validity Measure**						
Algorithm	**RI**	**ARI**	**SI**	**DBI**	**DI**	**CSI**	**ER (%)**
AMOC	0.6779	**0.2750**	**0.6341**	**0.8363**	**0.9790**	**1.4021**	**60.2804**
SOM	**0.6802**	0.1936	0.2248	1.2032	0.3321	2.0545	76.1682

Table T3 *Validity measures for Breast Cancer data of various algorithms*

BREASTCANCER	**Validity Measure**						
Algorithm	**RI**	**ARI**	**SI**	**DBI**	**DI**	**CSI**	**ER (%)**
AMOC	**0.9389**	**0.8771**	**0.7225**	0.7795	1.6440	**1.0531**	**3.1474**
SOM	0.9125	0.8230	0.5972	**0.7619**	**1.7615**	1.0952	4.5780

Table T4 *Validity measures for Half moon data of various algorithms*

HALF-MOON	**Validity Measure**						
Algorithm	**RI**	**ARI**	**SI**	**DBI**	**DI**	**CSI**	**ER (%)**
AMOC	**0.7955**	**0.5818**	**0.6348**	0.8003	2.3328	**1.0627**	**11.5252**
SOM	0.6219	0.2444	0.4920	**0.7786**	**2.3894**	1.0641	25.2011

Table T5 *Validity measures for Path Based data of various algorithms*

PATH BAESD	Validity Measure						
Algorithm	**RI**	**ARI**	**SI**	**DBI**	**DI**	**CSI**	**ER (%)**
AMOC	**0.7493**	**0.4642**	**0.7312**	0.678	2.5392	1.0144	**25.333**
SOM	0.7472	0.4606	0.5397	**0.6491**	**2.5957**	**0.9808**	25.6667

Table T6 *Validity measures for Spiral data of various algorithms*

SPIRAL	Validity Measure						
Algorithm	**RI**	**ARI**	**SI**	**DBI**	**DI**	**CSI**	**ER (%)**
AMOC	**0.6471**	**0.0608**	**0.4987**	**0.681**	2.1193	**0.8169**	83.9744
SOM	0.5547	-0.0047	0.3667	0.8704	**2.2019**	1.2457	**67.6282**

2. Validity measure for a given dataset and for each algorithm.

Table T7: *Number of cluster identified using AMOC*

Validity Index	Public Data Sets			Shape Sets		
	Iris(3)	**Glass(6)**	**BC(2)**	**HM(2)**	**PB(3)**	**Spiral(3)**
RI	**3**	8	**2**	**2**	7	10
	0.9495	0.7444	**0.9416**	**1**	0.8113	0.6639
ARI	**3**	8	**2**	**2**	6	10
	0.8857	0.2878	**0.8826**	**1**	0.5315	0.0842
SI	2	**6**	**2**	**2**	**3**	8
	0.9518	**0.6532**	**0.7164**	**0.5176**	**0.7297**	0.5275
DBI	2	**6**	**2**	**2**	**3**	8
	0.2319	**0.7542**	**0.7835**	**0.8997**	**0.651**	0.7257
DI	2	4	**2**	**2**	**3**	7
	6.4184	0.7174	**1.6302**	**2.0362**	**2.5392**	1.9861
CSI	2	3	**2**	**2**	**3**	8
	0.4025	2.4288	**1.0602**	**1.2254**	**0.9837**	1.0641
ER%	**3**	8	**2**	**2**	7	10
	0.9495	0.7444	**0.9416**	**1**	0.8113	0.6639

Table T8: *Number of cluster identified using SOM*

Validity Index	Public Data Sets			Shape Sets		
	Iris(3)	**Glass(6)**	**BC(2)**	**HM(2)**	**PB(3)**	**Spiral(3)**
RI	4	10	**2**	**2**	9	9
	0.9077	.7445	**0.9125**	**0.622**	0.7908	6.559
ARI	4	7	**2**	6	**3**	9
	0.7813	.2777	**0.823**	0.3152	**0.4607**	0.0718
SI	**3**	3	**2**	**2**	**3**	**3**
	0.6394	.5895	**0.5973**	**0.492**	**0.5398**	**0.3667**
DBI	2	7	**2**	6	6	9
	0.3827	.7325	**0.7619**	0.5637	0.6432	0.7666
DI	2	3	**2**	**2**	2	**3**
	2.613	1.1613	**1.7615**	**2.3894**	2.642	**2.2019**
CSI	2	**6**	**2**	**2**	**3**	6
	0.5106	**2.0545**	**1.0952**	**0.8181**	**0.9808**	1.0277
ER%	4	10	**2**	**2**	9	9
	0.9077	.7445	**0.9125**	**0.622**	0.7908	6.559