Hard C-means Clustering Algorithm in Gene Expression Data

Arabinda Panda[1] & Satchidananda Dehuri[2]

[1]*Department of Computer Science & Engineering,*
Modern Engineering & Management Studies, Balasore, Odisha
[2]*Department of Systems Engineering, Ajou University, San 5, Woncheon-dong,*
Republic of Korea
E-mail: arbind.omm@gmail.com[1]

Abstract – Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorial, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur. In this paper we describe hard C-means clustering algorithm in different biological data.

I. Introduction

The Hard clustering algorithm allocates each pattern (or, data point) to a single cluster during its operation. The Hard C-means clustering (HCM) algorithm is one of the best-known squared error- based clustering algorithm [7], [8], [9], [10]. It is very simple and can be easily implemented in solving many practical problems. It can work very well for compact and hyper-spherical clusters. The time complexity of Hard C-means is '$O(n \times C \times N)$' and space complexity is '$O(n + C)$', where 'n' is number of data points, 'N' is number of feature and 'C' is number of cluster in consideration. Since 'C' and 'N' are usually much less than 'n', Hard C-means can be used to cluster large data sets in least time. The HCM algorithm has been extensively applied in several areas [1], [2], [3], [4]. HCM algorithm has been extensively studied in the past for its applicability to pattern recognition and machine learning data.

II. Hard C-means (HCM) Clustering Algorithm

This section describes HCM clustering algorithm and their application to machine intelligence data as well as to bioinformatics data. First, basic steps involved in HCM clustering algorithm has been explained in details and then experiment are carried out to assess the performance of HCM clustering algorithm.

The HCM clustering algorithm ([1], [2], [3], [4]), one of the most widely used clustering techniques, attempts to solve the clustering problem by optimizing a objective function J, which is Mean Square Error (MSE) of formed cluster. The objective function J is given as follows:

$$Minimize(J) = \sum_{j=1}^{c} \sum_{x_i \in v_j} \left\| x_i - v_j \right\|^2$$

j=1, 2, …,c and i = {1, 2, .. .,n}.

The basic steps involved in HCM clustering algorithm are briefly described in Table 2.1.

Table 2.1: Hard C-means Clustering Algorithm

Notations:

X: Datasets;

X_i : i^{th} Data point of X;

x: any data point in X;

n : Total Number of data point in X; N : Cluster Center after updation;

Dimension of each data;

V_j = j^{th} Cluster;

C : Number of Cluster

v_j = j^{th} Cluster Center;

v^*_j = jth Cluster Center after Updation

Output:

$V_1 , V_2 ,...V_C$.

Objective Function:

$$J = \sum_{j=1}^{C} \sum_{x_i \in v_j} \left\| x_i - v_j \right\|^2 \; j \in 1, 2, ...C \text{ and}$$

i= { 1, 2, …n }

Hard C-means Clustering Algorithm:

Step 1: Chose "*C*" initial cluster centers (i.e. Prototype vector) $v_1 , v_2 , . . . v_C$ either randomly or using any intelligent techniques from the given 'n' data points $X_1 , X_2 , . . . , X_n$.

Step 2: Now compute $\left\| X_i - v_j \right\|$ *and* $\left\| X_i - v_p \right\|$ *f or* $p \in 1, 2, . . .C,$ *but* $p \# j$.

if $\left\| X_i - v_j \right\| < \left\| X_i - v_p \right\|$ *f or* $p \in 1, 2, . . .C,$ *but* $p \# j$ where X_i , $i = 1, 2, , n$;

Then put data X in cluster V_j.

If any ties occurred, then resolved it arbitrarily.

Step 3: Compute new cluster centers $\left\{ v_1^*, v_2^*, ..., v_c^* \right\}$ as follows:

$$v_j^* = \frac{1}{\left| v_j \right|} \sum_{x_i \in v_j} x_i \text{ , Where j = 1, 2, 3…,C}$$

and $\left| v_j \right|$ = The number of elements belonging to cluster v_j

Step 4: If $v_j^* = v_j$; $j = 1, 2, . . . ,C$; then terminate. Otherwise repeat from step 2.

Step 5: If the process does not terminate at Step 4 normally, then it is executed for a maximum "fixed number of iterations".

To validate the feasibility and performance of the HCM clustering algorithm, HCM clustering has been implemented in MATLAB 7.0 (Intel C2D processor, 2.0 GHz, *2GB* RAM) and applied it to Machine learning data as well as to bioinformatics data. Since the datasets used for simulation studies possess class label information, *clust ering accuracy* has been used as cluster validation metric to judge quality of the cluster formation algorithm.

Complete result of Hard C-means clustering for machine learning and bioinformatics data has been given in appendix C and appendix E. Appendix C contains the details of data distribution after simulation of HCM clustering algorithm and Appendix E contains the details of data point wrongly clustered in each cluster after simulation of HCM clustering algorithm. Here in this section, result obtained for HCM clustering algorithm has been discussed in brief.

Table 2.2 represents the result of Iris data. Overall accuracy has been achieved up to 88.67%. The total count error in this case is 17. It may be noted that 100% accuracy has been obtained for cluster 1. Similar sort of result for Iris data has been also reported in literature [5], [6], [7].

Table 2.3 represents the result of WBCD data. 95.75% accuracy has been achieved in this case.

Table 2.4 to Table 2.7 represents the result of subtypes of Breast data. Maximum ac- curacy was obtained for Breast Multi data A (79.61%) whereas the least accuracy for Breast data B was (53.06%). The reason of the less accuracy could be probably Breast data B is more overlapping in nature and is having nonlinear structure.

Table 2.8 to Table 2.11 represents the result of subtypes of DLBCL data (Diffused Large B-cell Lymphoma). DLBCL D is of highly overlapping in nature and that's why least accuracy of 42.64% has been obtained in this case. The data of DLBCL B is of highly distinctively separated in nature compared to other data such as DLBCL (A, C, D) and that is the reason higher accuracy was obtained in case of DLBCL B.

Table 2.12 represents the result of Lung Cancer. Accuracy up to 72.08% has been obtained. In this, cluster 2 and cluster 4 are highly separable in nature compared to cluster 1 and cluster 3. 95% accuracy was obtained for cluster 2 and cluster 4, whereas least accuracy was obtained for cluster 3 i.e., 47%.

Table 2.13 represents the result of St. Jude Leukemia data. For this data accuracy up to 85.08% was obtained. The data in this case is of highly separable in nature, 100% accuracy has been obtained for cluster 2 and least one was obtained for cluster 5.

Table 2.2: Result of Iris Data (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	50	50	50	150
The number of data point wrongly clustered	0	4	13	17
The number of data point correctly clustered	50	46	37	133
Accuracy (%)	100	92	74	88.67

Table 2.3: Result of WBCD Data (Hard C-means)

	Cluster 1	Cluster 2	Total
The right number of data point	444	239	683
The number of data point wrongly clustered	9	20	29
The number of data point correctly clustered	435	219	654
Accuracy (%)	97.97	91.63	95.75

Table 2.4: Result of Breast data A (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	11	51	36	98
Number of data point wrongly clustered	1	22	4	27
Number of data point correctly clustered	10	29	32	71
Accuracy (%)	90.91	56.86	88.89	72.44

Table 2.5: Result of Breast data B (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	12	11	7	19	49
Number of data point wrongly clustered	0	5	5	13	23
Number of data point correctly clustered	12	6	2	6	26
Accuracy (%)	100	54.54	28.57	31.58	53.06

Table 2.6: Result of Breast Multi data A (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	26	26	28	23	103
The number of data point wrongly clustered	2	1	18	0	21
The number of data point correctly clustered	24	25	10	23	82
Accuracy (%)	92.31	96.15	35.71	100	79.61

Table 2.7: Result of Breast Multi data B (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	5	9	7	11	32
The number of data point wrongly clustered	3	2	3	7	15
The number of data point correctly clustered	2	7	4	4	17
Accuracy (%)	40	77.78	57.14	36.36	53.13

Table 2.8: Result of DLBCL A (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	49	50	42	141
The number of data point wrongly clustered	26	22	18	66
The number of data point correctly clustered	23	28	24	75
Accuracy (%)	46.94	56	57.14	53.19

Table 2.9: Result of DLBCL B (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	42	51	87	180
The number of data point wrongly clustered	18	7	15	40
The number of data point correctly clustered	24	44	72	140
Accuracy (%)	57.14	86.27	82.76	77.78

Table 2.10: Result of DLBCL C (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	17	16	13	12	58
Number of data point wrongly clustered	1	9	12	6	28
Number of data point correctly clustered	16	7	1	6	30
Accuracy (%)	94.11	43.75	7.69	50	51.72

Table 2.11: Result of DLBCL D (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	19	37	24	49	129
The number of data point wrongly clustered	13	28	13	20	74
The number of data point correctly clustered	6	9	11	29	55
Accuracy (%)	31.58	24.32	45.83	59.18	42.64

Table 2.12: Result of Lung Cancer (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	139	17	21	20	197
The number of data point wrongly clustered	42	1	11	1	55
The number of data point correctly clustered	97	16	10	19	142
Accuracy (%)	69.78	94.11	47.62	95	72.08

Table 2.13: Result of St. Jude Leukemia data (Hard C-means)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Total
The right number of data point	15	27	64	20	43	79	248
The number of data point wrongly clustered	15	0	3	4	14	1	37
The number of data point correctly clustered	0	27	61	16	29	78	211
Accuracy (%)	0	100	95.31	80	67.44	98.73	85.08

III. Conclusion

There is no efficient and universal method for identifying the initial partitions in Hard C-means clustering algorithm. The convergence centroids vary with different initial points and that may results in suboptimal solution. This particular limitation of Hard C-means clustering algorithm has been extensively studied and this problem is still an open issue of research.

IV. Reference

[1] N. Ye, The Hand Book of Data Mining. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2003.

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann, Elsevier, 2006.

[3] R. Dubes and A. Jain, Algorithms for Clustering Data. Prentice Hall, 1988. [10] A. K. Jain, M. N. Murty, and P. J. Flynn, Data clustering: A Review, ser. 3. ACM Computing Surveys, September 1999, vol. 31.

[4] N. R. Pal, J. C. Bedzek, and E. C. K. Taso, “Generalized clustering networks and kohonen’s self-organizing scheme,” IEEE Trans. on Neural Networks, vol. 3, no. 4, pp. 546–557, July 1993.

[5] A. I. Gonzalez, M. Graiia, and A. D’Anjou, “An analysis of the GLVQ algorithm,” IEEE Trans. on Neural Networks, vol. 6, no. 4, pp. 1012–1016, July 1995.

[6] N. B. Karayiannis, J. C. Bezdek, N. R. Pal, R. J. Hathaway, and P. I. Pai, “Repairs to GLVQ: A new family of competitive learning schemes,” IEEE Trans. on Neural Networks, vol. 7, no. 5, pp. 1062–1071, Sept. 1996.