ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Effective way of Clustering Large Data Sets using Multi-Level Data Processing

N. SreeRam

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India, sriramnimmagadda@gmail.com

Abstract.

Cluster is a gathering of items that belongs with a similar class. At the end of the day, comparative articles are assembled in one cluster and different objects are gathered in another cluster. clustering is the way toward making a group of conceptual articles into classes of comparable items. Clustering analysis is comprehensively utilized as a part of numerous applications, for example, statistical surveying, design acknowledgment, information investigation, and picture handling. Clustering can likewise help advertisers find gatherings in their client base. Also, they can portray their client bunches in view of the buying designs. Clustering is a popular strategy for implementing parallel processing applications because it enables companies to leverage the investment already made in PCs and workstations. In addition, it's relatively easy to add new CPUs simply by adding a new PC to the network. In this paper we study k-mode clustering and implementing those for date classification.

Keywords: clustering, parallel processing, client bunch, work stations.

1. Introduction

The k-modes clustering algorithm is an extension to the standard k-means clustering algorithm for clustering categorical data. In data mining, k-means is the mostly used algorithm for clustering data because of its efficiency in clustering very large data. However, the standard k-means clustering process cannot be applied to categorical data due to the Euclidean distance function and use of means to represent cluster centres. To use k-means to cluster categorical data, we convert each unique category to a dummy binary attribute and used 0 or 1 to indicate the categorical value either absent or present in a data record. This approach is not suitable for high dimensional categorical data. The k-modes approach

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

modifies the standard k-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centres and updating modes with the most frequent categorical values in each of iterations of the clustering process. These modifications guarantee that the clustering process converges to a local minimal result and the efficiency of the clustering process is maintained.

To cluster a categorical data set X into k clusters, the k-modes clustering process consists of the following steps:

Step 1: Randomly select k unique objects as the initial cluster centres (modes).

Step 2: Calculate the distances between each object and the cluster mode; assign the object to the cluster whose centre has the shortest distance to the object; repeat this step until all objects are assigned to clusters.

Step 3: Select a new mode for each cluster and compare it with the previous mode. If different, go back to Step 2; otherwise, stop.

This clustering process minimises the following k-modes objective function

$$F(U,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}u_{i,l}d(x_{i,j},z_{l,j})$$

2. Literature Review

The two different approaches that fall under this are: top-down and bottom-up. The most important difference between k-mode and k-means clustering is he hierarchy. The difference lies from how the algorithm can be implemented o how the results can be interpreted.

Top-down hierarchical clustering:

The data is divided into 2 clusters (using k-means with k=2k=2, for example). Then, for each cluster, the process is repeated, until all the clusters are too small or too similar for further clustering, or until a preset number of clusters is reached.

Bottom-up hierarchical clustering:

Each data item consists of its own cluster. We then look for the two items that are most similar, and combine them in a larger cluster. The process is repeated until all the clusters that are left are too dissimilar to be gathered together, or until a preset number of clusters is reached.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

In k-means clustering, the main objective is to divide the data into k sets simultaneously. A better approach is to take k items from data set as initial representatives for the cluster, assign all items to the cluster whose representative is closest, and then calculate the cluster mean which becomes the new representative, until all clusters stay the same.

The hierarchical clustering methods can further be divided based on the measure of similarity:

Single-link clustering:

It considers the distance between two clusters as the minimum distance from any member of one cluster to any member of the other. If similarities are found, the similarity between a pair of clusters is considered to be the greatest similarity from any member of one cluster to any member of the other. It is also known as the minimum method.

Complete-link clustering

It considers the distance between two clusters as the longest distance from any member of one cluster to any member of the other. It is also known as the maximum method

Average-link clustering

It considers the distance between two clusters as the average distance from any member of one cluster to any member of the other. It is also known as the minimum variance method

3. Implementation

The given data is clustered by the k-mode method which aims to split the items into n groups such that the distance between the item and the assigned cluster modes is minimized. Conventionally, simple-matching distance is used for determination of the dissimilarity between two items or clusters. The computation takes place by keeping a count of the number of mismatches in the variables encountered. As an alternative, the weighted version of this distance is calculated by the frequencies of the categorical data. If a preset matrix is provided, it might be possible that none of the objects will be closest to one or more modes. In this case few clusters than supplied modes will be returned, intimating a warning.

The dataset used (iris species) for the analysis contains 200 data points, further divided into 153 training sets and 47 testing samples, and 4 attributes for each sample which are utilized for classification. Based on the above algorithm, the data has been divided into 3 clusters.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Using R-studio:

kmode(data, modes, itrn.max = 10, weight = FALSE)

Arguments

data

Categorical data organised in a matrix, with objects in rows and variables in columns, or data frames.

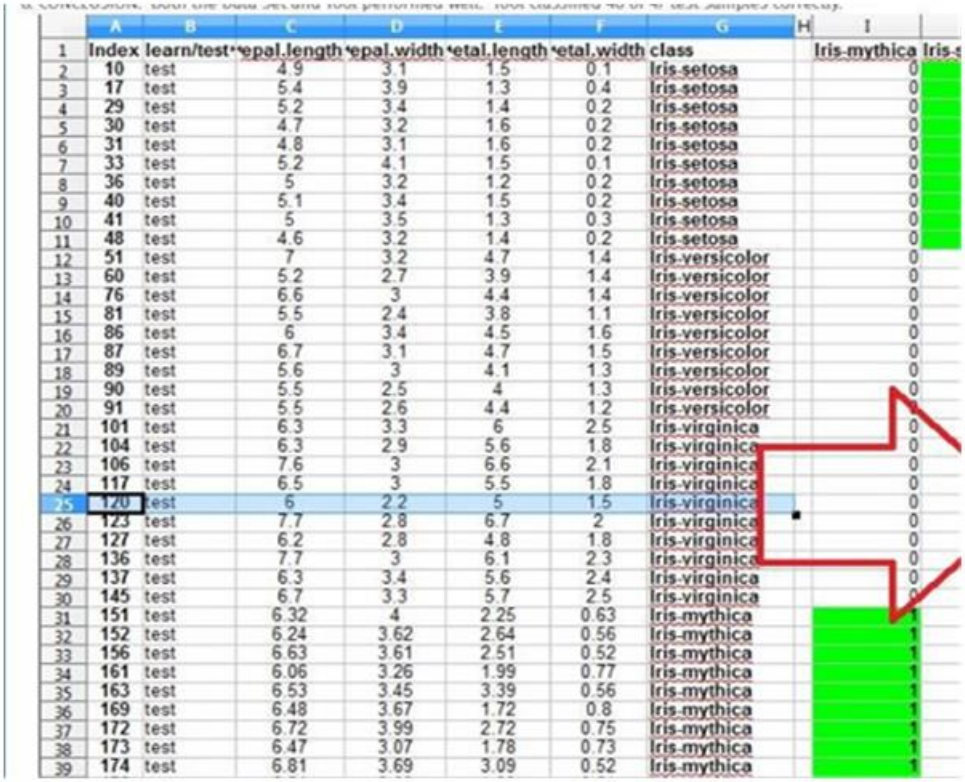
modes

Either the number of modes or the preset number of cluster modes. A random set of distinct rows in data is assigned as the initial modes.

itrn.max : maximum number of iterations allowed.

Weighted

Simple-matching distance between objects, either weighted or unweighted, is used



	A	B	C	D	E	F	G	H	I	
1	Index	learn/test	epal.length	epal.width	etal.length	etal.width	class		Iris-mythica	Iris-s
2	10	test	4.9	3.1	1.5	0.1	Iris-setosa		0	
3	17	test	5.4	3.9	1.3	0.4	Iris-setosa		0	
4	29	test	5.2	3.4	1.4	0.2	Iris-setosa		0	
5	30	test	4.7	3.2	1.6	0.2	Iris-setosa		0	
6	31	test	4.8	3.1	1.6	0.2	Iris-setosa		0	
7	33	test	5.2	4.1	1.5	0.1	Iris-setosa		0	
8	36	test	5	3.2	1.2	0.2	Iris-setosa		0	
9	40	test	5.1	3.4	1.5	0.2	Iris-setosa		0	
10	41	test	5	3.5	1.3	0.3	Iris-setosa		0	
11	48	test	4.6	3.2	1.4	0.2	Iris-setosa		0	
12	51	test	7	3.2	4.7	1.4	Iris-versicolor		0	
13	60	test	5.2	2.7	3.9	1.4	Iris-versicolor		0	
14	76	test	6.6	3	4.4	1.4	Iris-versicolor		0	
15	81	test	5.5	2.4	3.8	1.1	Iris-versicolor		0	
16	86	test	6	3.4	4.5	1.6	Iris-versicolor		0	
17	87	test	6.7	3.1	4.7	1.5	Iris-versicolor		0	
18	89	test	5.6	3	4.1	1.3	Iris-versicolor		0	
19	90	test	5.5	2.5	4	1.3	Iris-versicolor		0	
20	91	test	5.5	2.6	4.4	1.2	Iris-versicolor		0	
21	101	test	6.3	3.3	6	2.5	Iris-virginica		0	
22	104	test	6.3	2.9	5.6	1.8	Iris-virginica		0	
23	106	test	7.6	3	6.6	2.1	Iris-virginica		0	
24	117	test	6.5	3	5.5	1.8	Iris-virginica		0	
25	120	test	6	2.2	5	1.5	Iris-virginica		0	
26	123	test	7.7	2.8	6.7	2	Iris-virginica		0	
27	127	test	6.2	2.8	4.8	1.8	Iris-virginica		0	
28	136	test	7.7	3	6.1	2.3	Iris-virginica		0	
29	137	test	6.3	3.4	5.6	2.4	Iris-virginica		0	
30	145	test	6.7	3.3	5.7	2.5	Iris-virginica		0	
31	151	test	6.32	4	2.25	0.63	Iris-mythica		1	
32	152	test	6.24	3.62	2.64	0.56	Iris-mythica		1	
33	156	test	6.63	3.61	2.51	0.52	Iris-mythica		1	
34	161	test	6.06	3.26	1.99	0.77	Iris-mythica		1	
35	163	test	6.53	3.45	3.39	0.56	Iris-mythica		1	
36	169	test	6.48	3.67	1.72	0.8	Iris-mythica		1	
37	172	test	6.72	3.99	2.72	0.75	Iris-mythica		1	
38	173	test	6.47	3.07	1.78	0.73	Iris-mythica		1	
39	174	test	6.81	3.69	3.09	0.52	Iris-mythica		1	

Figure.1: Dataset

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

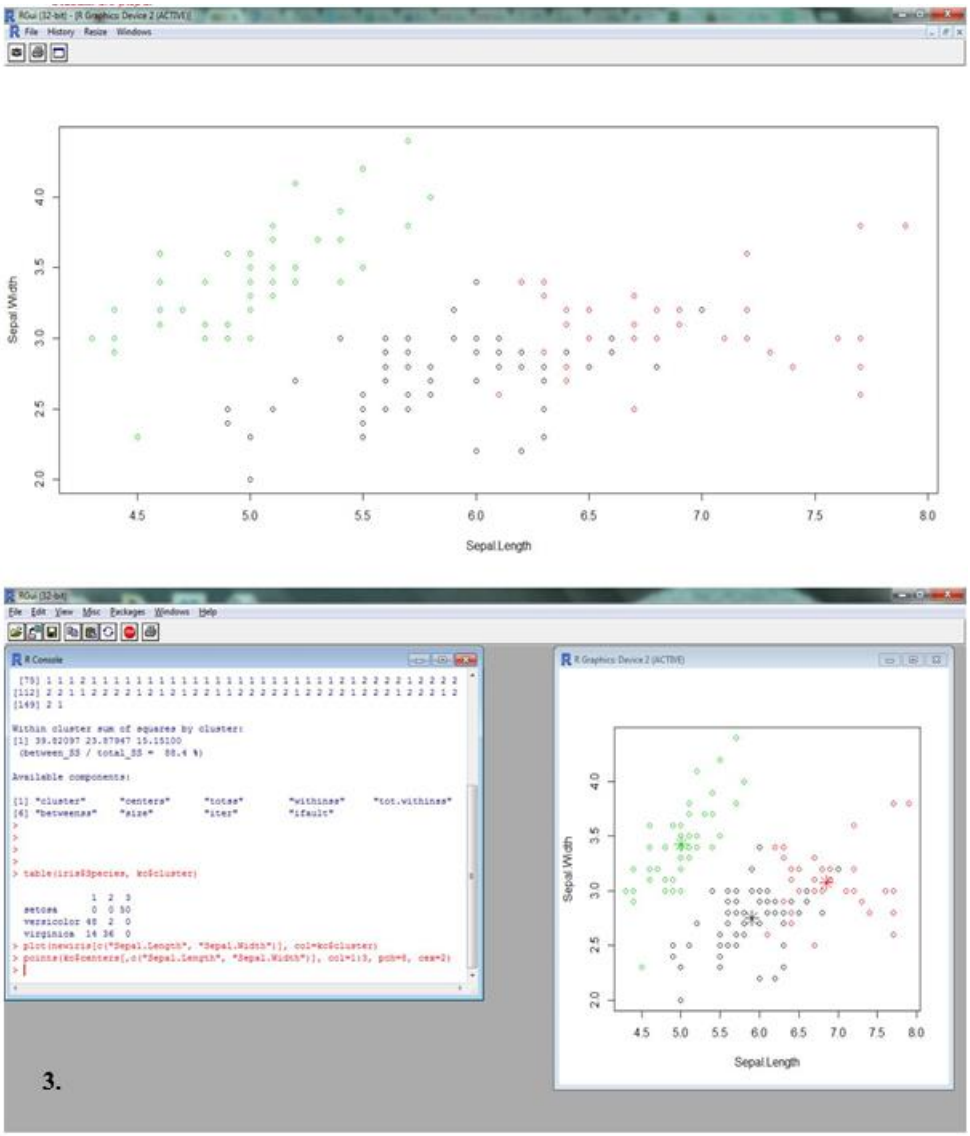


Figure 2: Results of clustering

4. Conclusion

Determination of grouping in a set of unlabelled information on the basis of its features is the main objective of clustering. There is no absolute "best" criterion which would be independent of the final aim of the clustering. Subsequently, it is the user who must supply the standards, in such a way that the outcome suits suit their requirements.

The above data set has been classified for the categorical data using k-mode clustering data. From this data implementation we got three clusters based on mode values. This paper presents the analysis of k-mode clustering, highlighting its advantages and limitation.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

References:

[1]Anderberg, M.R. 1973. Cluster Analysis for Applications. Academic Press.

[2]Ball, G.H. and Hall, D.J. 1967. A clustering technique for summarizing multivariate data. Behavioral Science, 12:153–155.

[3]Bobrowski, L. and Bezdek, J.C. 1991. c-Means clustering with the l 1 and l ∞ norms. IEEE Transactions on Systems, Man and Cybernetics, 21(3):545–554.

[4]Cormack, R.M. 1971. A review of classification. J. Roy. Statist. Soc. Serie A, 134:321–367.

[5]Dubes, R. 1987. How many clusters are best? An experiment. Pattern Recognition, 20(6): 645–663.

[6]Ester, M., Kriegel, H.P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA: AAAI Press, pp. 226–231.

[7]Guo Tao, Ding Xingu, Li Yefeng, Parallel k-modes. Algorithm based on MapReduce.

[8]S.Aranganayagi, K.ThangaveI, S.Sujatha, New Distance.Measure based on the Domain for Categorical Data.

[9]Guo Tao, Ding Xingu, Li Yefeng, Parallel k-modes.Algorithm based on MapReduce.

[10]https://www.aaai.org/Papers/IJCAI/2007/IJCAI07-447.pdf

[11]https://shapeofdata.wordpress.com/2014/03/04/k-modes/

[12]http://www.daylight.com/meetings/mug04/Bradshaw/why_k-modes.html