A Simple Approach to Clustering in Excel

Aravind H
Center for Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India

C Rajgopal
Center for Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India

K P Soman
Center for Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India

ABSTRACT
Data clustering refers to the method of grouping data into different groups depending on their characteristics. This grouping brings an order in the data and hence further processing on this data is made easier. This paper explains the clustering process using the simplest of clustering algorithms - the K-Means. The novelty of the paper comes from the fact that it shows a way to perform clustering in Microsoft Excel 2007 without using macros, through the innovative use of what-if analysis. The paper also shows that, image processing operations can be done in excel and all operations except displaying an image do not require a macro. The paper gives a solution to the problem of reading an image in excel by introducing a user defined add-in. The paper also has explained and implemented image segmentation as an application of clustering. This paper aims at showing that Microsoft Excel is a great tool as far as technical learning is concerned for the fact that, it can implement almost all algorithms and processes, and is very successful in providing the first hand exposure to an novice student.

General Terms
Machine learning, Clustering, Image processing.

Keywords
Data clustering, K-Means, Image Segmentation, Excel add-in to read image, Microsoft Excel

1. INTRODUCTION
It is of no doubt that, information is the driving force of the world. But what is information: To be exact, it is a collection of meaningful data. Each day in the world of computing is manipulating on billions of data to extract some information from them. The data acts like an ideal gas literally, and they tend to fill the best and largest storage in no time. Hence managing data is a complex job. Grouping them into different groups as soon as they are obtained will bring an order in the data .This will help in reducing the computation complexity required in further processing and managing it. The word "Data clustering" refers to the process of partitioning a set of data into a set of meaningful sub classes called clusters. Data clustering has immense number of applications in every field of life. One important application of clustering is in the field of data mining. Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. For example, in case of detection of diseases like tumors, the scanned pictures or the x-rays is subjected to hierarchical clustering. Here clusters are formed using a variety of images available for a specific part of the body along with valid records. Such clusters are created for all body parts. Now the tumor affected part in the body is located by comparing the test image with the images in these clusters. Once the body part is located, image is sent to that specific matching cluster and matched with all the images in that particular cluster. Now the image, with which the query image has the most similarities, is retrieved and the record associated to that image is taken. Using this technique really fine tumor can be detected [1]. By using clustering an enormous amount of time in finding the exact match from the database is reduced.

In the management side, a normal requirement of predicting the sales of a product at different cities is achieved by clustering demographically similar cities. Another application of clustering is load balancing in application servers. Load balancing is an enterprise-level feature in which the application server automatically alternates requests among the server instances in a cluster. Clustering enables application servers to route requests to a running server instance when the original server instance goes down [1]. Clustering is also used to improve the performance (i.e. perplexity) of language models as well as to compress language models [2].

Here we have explained on how to implement one of the simplest of the clustering algorithms, the K-Means. This is done with the help of two add-in packages available with Microsoft Excel - the Solver and What-If analysis. As an application of clustering, the details regarding, how segmentation of a picture using clustering can be implemented in Microsoft Excel is explained with this paper.

1.1 Microsoft Excel
Microsoft Excel is a basic learning tool of great potential. Excel's forte is performing numerical calculations, organize data, compare as well present data graphically [8]. Using Excel it is possible to implement almost all algorithms and thus helps the learner to get an exposure about the functionality of the algorithm. To an extend it is true to say that Excel is an intelligent software except that, it is only as intelligent as the user. Excel package provides with it a large amount of features, most of them as add-in packages. In our paper we are mainly concerned about two of the add-in packages – the "Solver" and the "What-If Analysis". We have also utilized the conditional formatting to provide a look and feel to the display of output.

"Solver" is basically an optimization problem solver which takes an objective function and a set of constraints and provides back

an optimum solution i.e. best values for the variables by which the objective function is made of so that the objective function is minimized or maximized(which is user requirement). The "best" or optimal solution may mean maximizing profits, minimizing costs, or achieving the best possible quality.

"What-If Analysis" feature helps in executing the same set of operations for different inputs and then records or prints the outputs for each of those input sets. This makes life easier by reducing the workload of doing the same set of operations for different input sets manually. The implications of What-If analysis were understood from [5], [6].

Conditional Formatting is a method by which specific operation like highlight interesting cells or ranges of cells, emphasize unusual values, visualize data by using data bars, color scales, and icon set, doing some mathematical manipulation etc are easily implemented, depending on the rule set defined for that specific set of cells. This rule sets will be active all the time i.e. When cell values are changed the rules are reapplied.

2. CLUSTERING ALGORITHMS

The performances given by clustering algorithms are heavily dependent on the spread of the data and for this reason there are more than one clustering algorithms which are developed over time. Some of them include the K-Means, K-Medoids, the EM algorithm, different types of linkage methods, the mean-shift algorithm, algorithms that minimize some graph-cut criteria etc. Thus it is true to say that, a universal clustering algorithm remains an elusive goal.

2.1 K-means Clustering

K-means clustering is one of the basic clustering algorithms in the machine learning domain. The inference of this algorithm is based on the value of 'k' which is the number of clusters that can be found in an n-dimensional dataset. Usually the value of 'k' is assumed or known a-priori. In k-means algorithm, since it is considered that there is 'k' number of clusters; we consider that there are 'k' number of cluster means (cluster centers), where the cluster mean is the average of all the data-points falling under each cluster. The end result of the k-means clustering algorithm is that each data point in the data-set is grouped into 'k' clusters around the 'k' cluster means. If the data points are tightly surrounding the cluster means, then it is considered as a highly cohesive and good cluster, else it is not. Hence the cluster compactness forms the metric of quality of the k-means algorithm. Thus as per [7] we can define a measure of cluster compactness as the total distance of each data point of a cluster from the cluster mean which is given by,

$$\sum_{x_i \hat{I} C_k} \left\| X_i - \bar{X}_k \right\|^2 = \sum_{i=1}^{m} Z_{ki} \left\| X_i - \bar{X}_k \right\|^2$$

where the cluster mean is defined as $\bar{X}_k = \frac{1}{m_k} \sum_{x_i \hat{I} C_k} X_i$

and $m_k = \sum_{i=1}^{m} Z_{ki}$ is the total number of points allocated to cluster k. The parameter Z_{ki} is an indicator variable indicating the suitability of the i^{th} data point X_i to be a part of the k^{th} cluster. The suitability is determined by considering points at a minimum distance to the cluster mean to be a part of the cluster. The value of the indicator variable can be considered to be 1 when the i^{th} data point falls in the k^{th} cluster and for the other situations as 0.Again from [7], the total goodness of the clustering will then be based on the sum of the cluster compactness measures for each of the 'k' clusters. Using the indicator variables Z_{ki} then we can define the overall cluster goodness as:

$$\varepsilon_K = \sum_{i=1}^{m} \sum_{k=1}^{K} Z_{ki} \left\| X_i - \bar{X}_k \right\|^2$$

Now the objective of the algorithm is to find an optimum $\bar{X}_k$ for the above equation so that the value of ε_K (the measure of overall cluster quality) reaches its minimum.

2.1.1 The Simplified Algorithm

The simplified algorithm for performing the k-means clustering can be given in the form of an iterative minimization of the overall measure of cluster quality $\mathcal{E}_K$. It can be elaborated in the following steps:

1. Given the value of 'k', choose 'k' arbitrary cluster means $\bar{x}_k$ and find all indicator parameters z_{ki} for each of the cluster means. This is essentially done by putting z_{ki} as '1' if the i^{th} data point has minimum distance to the k^{th} cluster mean compared to the distance from all the other cluster means and '0' if it is not at minimum distance.
2. Calculate the overall metric
$$\mathcal{E}_K = \sum_{i=1}^{m} \sum_{k=1}^{K} z_{ki} \left\| x_i - \bar{x}_k \right\|^2$$
3. Minimize the overall metric by assuming a new set of cluster means.
4. For the new set of cluster means calculate the new indicator parameter values as described in Step 1.
5. For the new set of cluster means and indicator parameter values, recalculate the new overall metric.
6. Repeat Steps 3 to 5 until $\mathcal{E}_K$ converges.

2.2 Implementation Details for K-Means Clustering

Clustering, as the name says is the process of grouping of data. In our case we take a group of data point in two dimensions. Let's assume data two dimensional data points as (3,13), (3, 10), (8,7), (5,2), (5,4), (2,11), (5,4),(2,14), (6,2),(4,2), (10,8), (8,9), (10,9), (10,12).

	A	B	C	D	E
5		TEST DATA			
6		SL NO	X	Y	
7		1	4	12	
8		2	5	10	
9		3	8	7	
10		4	5	3	
11		5	5	4	
12		6	2	11	
13		7	5	4	
14		8	3	8	
15		9	6	2	
16		10	7	4	
17		11	10	8	
18		12	8	9	
19		13	10	9	
20		14	10	12	

Figure 1. Data points

The data points are assumed here suitability such that the three distinct clusters can be identified easily (even visually). It must be kept in mind that the real time data points will not always provide such a visual advantage. Clustering can be done on any N dimensional data and for any data groups with any sought of compactness.

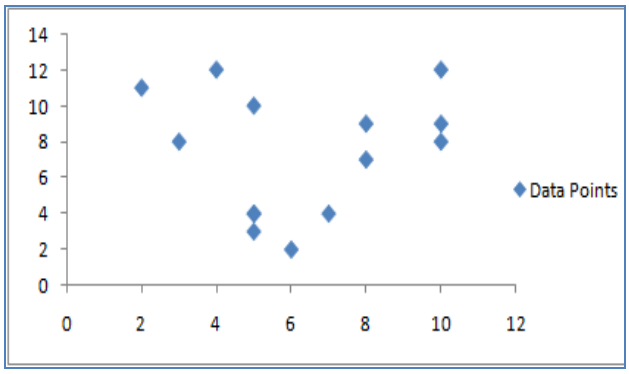

Figure 2. XY Plot of Data points

The K –Means algorithm starts with the initial assumption of K. Here we have assumed K as 3. Since K = 3, we have 3 cluster centers and they are initialized with random points. Lets the initial points assumed be (4, 4), (5, 12), (10, 6). This is entered in the separate cells in Microsoft excel sheet as shown in the figure 3 below.

	F	G	H	I	J	K
6	Cluster Center 1		Cluster Center 2		Cluster Center 3	
7	X	Y	X	Y	X	Y
8	4	4	5	12	10	6

Figure 3. Cluster Center's (Initially assumed)

Figure 4 is a plot of initial cluster centers and the data points. This is obtained by selecting the range of the data points (C7:D20) and three cluster points separately (X=F8, Y=G8), (X=H8, Y=I8) and (X=J8, Y=K8) and then plotting it using XY plot available with excel.

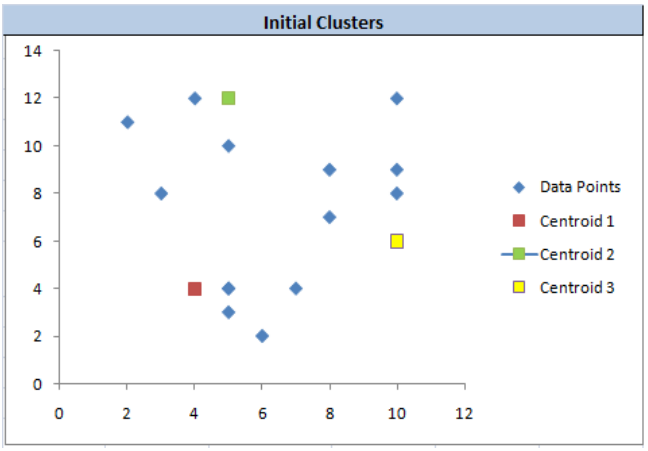

Figure 4. Cluster Center's (Initially assumed)

In K-Means algorithm the objective is to minimize the term $\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|X_i^{(j)} - C_j\right\|^2$, where X_i is the data point belonging to the cluster and C_j is the cluster center (centroid). For this reason, as the next step we need to find out the Euclidian distance of the each data point with all the centroids (cluster centers). Here, for easiness we utilize the "What-If Analysis" feature available with excel. As a prerequisite for "What-If Analysis" we need to make available a structure as shown below, where the X and Y values for the chosen index will be taken from the table shown in figure 1. The index value will be provided in the cell G11 and the corresponding X and Y values are obtained in the cells I11 and J11. The formula used in I11 is =INDEX(C7:C20,$G$11,1) and formula in J11 is =INDEX(D7:D20,$G$11,1).

	G	H	I	J
10	INDEX OF CURRENT POINT		X	Y
11	1		4	12

Figure 5. A requirement for what if analysis

We have 14 data points in our example and need to find the distance (Euclidian distance) of all the points with each of the assumed cluster centers. The below figure shows table structure designed to find out the different Euclidian distances using the "What-If Analysis" [5], [6].

	M	N	O	P	Q	R	S	T
6	INDEX	Distance to CENTROID 1	Distance to CENTROID 2	Distance to CENTROID 3	X	Y	CLASS	Minimum Distance
7	1	8	1	8.485281374	4	12	Cluster2	1
8	2	6.08276253	2	6.403124237	5	10	Cluster2	2
9	3	5	5.830951895	2.236067977	8	7	Cluster3	2.236067977
10	4	1.414213562	9	5.830951895	5	3	Cluster1	1.414213562
11	5	1	8	5.385164807	5	4	Cluster1	1
12	6	7.280109889	3.16227766	9.433981132	2	11	Cluster2	3.16227766
13	7	1	8	5.385164807	5	4	Cluster1	1
14	8	4.123105626	4.472135955	7.280109889	3	8	Cluster1	4.123105626
15	9	2.828427125	10.04987562	5.656854249	6	2	Cluster1	2.828427125
16	10	3	8.246211251	3.605551275	7	4	Cluster1	3
17	11	7.211102551	6.403124237	2	10	8	Cluster3	2
18	12	6.403124237	4.242640687	3.605551275	8	9	Cluster3	3.605551275
19	13	7.810249676	5.830951895	3	10	9	Cluster3	3
20	14	10	5	6	10	12	Cluster2	5
21							Sum of Minimum distances	28.45086703

Figure 6. Assignment of clusters

For "What-If Analysis", values 1 to 14 are written in leftmost column M (M7 to M20) as shown in the above screenshot of the table. In the first row of the second column(i.e. Cell N7), type the formula to find the Euclidian distance of centroid 1 from data point 1, i.e.=SQRT(($I$11-$F$8)^2+($J$11-$G$8)^2), when "INDEX OF CURRENT POINT" is set to 1, $F$8 $I$11 will refer to the, X coordinate of the data point 1 and J11will refer to the, Y coordinate of the data point 1. The $F$8 holds the X value of cluster center 1 and $G$8 holds the y value of cluster center 1.Similarly in the third column(i.e. cell O7) type the formula for finding distance of second centroid with the data point 1. i.e. =SQRT(($I$11-$H$8)^2+($J$11-$I$8)^2). In the next column (i.e. cell P7) do find the distance with the third centroid. i.e. =SQRT(($I$11-$J$8)^2+($J$11-$K$8)^2). In the last but one's column find out the class to which the data point 1 belongs. The formula used to obtain this is -=IF(MIN(N7:P7)=N7,"Cluster1",IF(MIN(N7:P7)=O7,"Cluster2", "Cluster3")), here MIN(N7:P7) takes the minimum distance from the distances calculated for the selected data point with the all the three cluster centers, 1 ,2 and 3. The formula also implements the functionality finding out from which column the minimum value comes. I.e. if the minimum comes from the second column we assign the point to "Cluster 1", if it comes from the third column we assign this point as a "Cluster 2" point and else if it comes from the fourth column, we assign "Cluster3". This is done in order to assign the point to a particular cluster. The "What-If Analysis " will do the above mentioned operations for all the point by changing the index value of the table shown in figure 5 from 1 to 14 and corresponding results are assigned to the cells from S7:S20.

Now we need to explicitly find out and keep the minimum distance of each point from among the three cluster centers. This calculation is what is being done in the last column. In the column labeled 'Minimum distance" we calculate the sum of minimum distances obtained for all the data point. Now we calculate the sum of the above calculated minimum distance values for all the points as shown in figure 6. The K-Means algorithm specifies to minimize the above mentioned sum of distances to the nearest cluster centers. The minimization is done with the help of "Solver" feature available with excel. A "Solver" basically solves an optimization problem (minimization or maximization problem) subjected to a set of constraints. The figure 7 depicts on, how to find optimal cluster centers using solver. As figure 7 show, the cell corresponding to the label "Set Target cells" is assigned with the excel cell address whose content value has to be minimized. Here it is $T$21 (objective function) and the cells corresponding to the label "By Changing Cells" should be provided with the address of cells whose value has to be changed/adjusted in order to achieve the minimization. Here it is $F$8:$K$8.

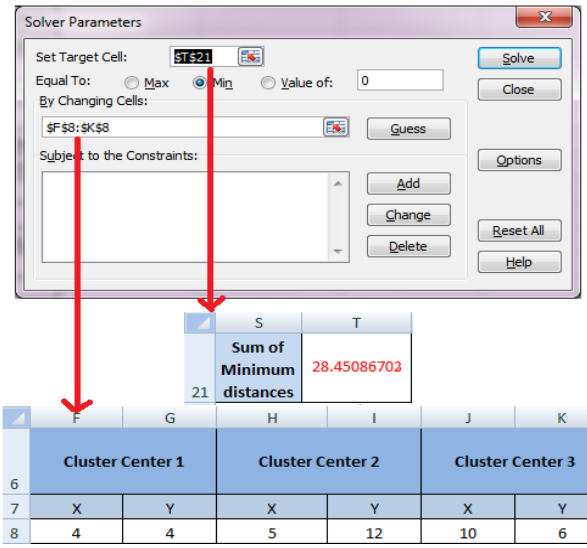

Figure 7. Assigning "Solver" inputs

On clicking the "Solve' button, the optimal values for all three centroids are obtained. The solver achieves this by incrementing or decrementing the values of the assumed cluster centers by small factor (this factor can be specified as a Solver parameter). Any change in the cluster center will reflect in the allocation of data points to the clusters. [This is a feature of What-If Analysis add-in]. After solving we can see that the cluster centers will come close to the center of corresponding clusters. The reallocation of data points can be seen in figure 9. Figure 8 shows the minimized distances value and optimal cluster centers.

Sum of Minimum distances	21.01111085

(a)

	F	G	H	I	J	K
6	Cluster Center 1		Cluster Center 2		Cluster Center 3	
7	X	Y	X	Y	X	Y
8	5.21815716	3.66736761	3.615385665	10.461533	9.47841197	8.75055345

(b)

Figure 8. (a) Minimized distance, (b) Optimal cluster centers

The new cluster centers obtained are (5.218, 3.667), (3.615, 10.461), (9.478, 8.750). The XY plot of the new cluster centers is shown in figure 10. From figure 10 it can be seen that the optimal cluster centers have moved more close to the center of each cluster.

	M	N	O	P	Q	R	S	T
6	INDEX	Distance to CENTROID 1	Distance to CENTROID 2	Distance to CENTROID 3	X	Y	CLASS	Minimum Distance
7	1	8.421202174	1.585811134	6.369612419	4	12	Cluster2	1.585811134
8	2	6.336387779	1.45951135	4.649445429	5	10	Cluster2	1.45951135
9	3	4.341093127	5.586329856	2.291322644	8	7	Cluster3	2.291322644
10	4	0.7021203	7.588918895	7.28869675	5	3	Cluster1	0.7021203
11	5	0.397788095	6.608223376	6.528705239	5	4	Cluster1	0.397788095
12	6	8.007746607	1.70276631	7.80940091	2	11	Cluster2	1.70276631
13	7	0.397788095	6.608223376	6.528705239	5	4	Cluster1	0.397788095
14	8	4.867433339	2.537294492	6.521750626	3	8	Cluster2	2.537294492
15	9	1.84157525	8.791131476	7.594035326	6	2	Cluster1	1.84157525
16	10	1.812626062	7.294317915	5.3582012	7	4	Cluster1	1.812626062
17	11	6.452730419	6.842694287	0.913990839	10	8	Cluster3	0.913990839
18	12	6.014616733	4.62178901	1.499314186	8	9	Cluster3	1.499314186
19	13	7.162610719	6.54976269	0.578161775	10	9	Cluster3	0.578161775
20	14	9.607225389	6.567356386	3.291040321	10	12	Cluster3	3.291040321
21							Sum of Minimum distances	21.01111085

Figure 9. Optimal clusters assignment

We can then plot the final points which are the optimal cluster centers as show in the figure 10. In figure 10 we can see that, three clusters are recognized. This is because the value of 'k' in our example is 3. It must be kept in mind that 'k' can be any higher value, and as 'k' increases the number of clusters which are recognized by the algorithm will also increase.

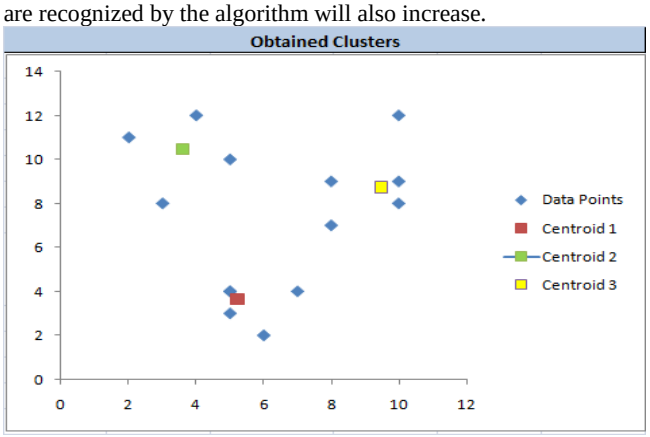

Figure 10. XY Plot of Data points and Optimal Cluster centers

2.3 An Application of Clustering

Image segmentation, an important problem in computer vision, is often formulated as a clustering problem. A color picture is made of three components R (Red), G (green) and B (components). As mentioned previously clustering groups the different similar data points depending on the Euclidian distance of its property values in property space. Here we have selected a color image and has clustered different colors in a picture. The number of clusters created will be dependent on the value of 'k'. Here we have chosen k = 4, since there are four colors in the considered image.

Due to the easiness to implement here we have chosen k-means algorithm for color component clustering. First of all we need to read the image to be clustered into excel sheet. This will give the corresponding pixel values(R, G and B) for the image. Reading of image is done well within excel by using a user defined add-in named "loadImageArray". The add-in is made available at [4]. The add-in was developed using C# [3].The add-in once installed can be obtain from the functions tab of excel as shown below.

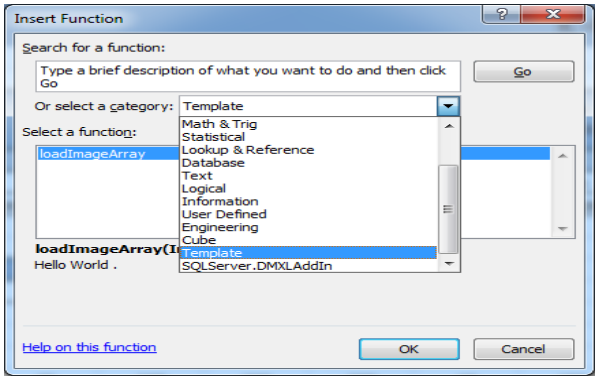

Figure 11. Image description of selecting "loadImageArray"

Once you select the template, we can see "loadImageArray" popping up in the function list. On selecting this function one can see the below given popup window. It will ask for four arguments – Location of image (should be given within double quotes), length of image, width of image and finally the color component you are interested in. i.e. Give 1 for acquiring red, 2 for acquiring green and 3 for acquiring blue pixels of the image.

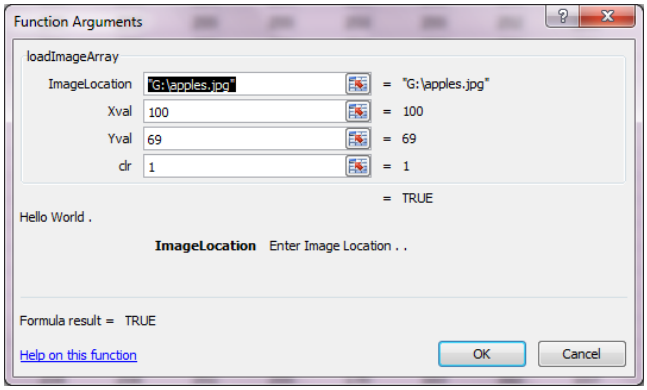

Figure 12. "loadImageArray" - details input screen

On clicking "ok", the first pixel will be shown in the selected cell of the excel sheet. Beginning from that cell, select a range 100 x 69 cells (for this example). Now select the formula in the starting cell where the pixel value was printed cell and press "Ctrl + Alt + Enter". Now we will get all the pixel value for red component. Do this same process to obtain the green and blue component by giving 2 and 3 in place of 1. As a second step we can plot this picture (Apples). This can be done by suitably coloring the corresponding cells of an excel sheet by the color we obtain from the pixel value combination(R, G and B which was obtained in three different sheets). We have achieved this coloring of excel

cells by using a small macro. Finally reduces the row height and column width of the excel cells to small values so that the image can be seen in the required size.

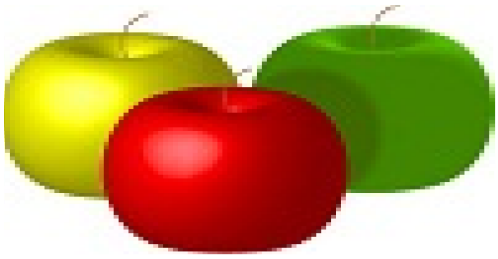

Figure 13. Apples (Input image for clustering)

We have the image pixel values for R, G and B components, read and stored in excel file in separate sheets. Now we can move to segmentation of pixel values. This segmentation is done using k-means clustering. Thus the first process, we need to do is to randomly select the cluster centers (centroids). The values selected should be within the range 0- 255 since pixel values can only be in this range. Here we have selected four cluster centers since we are interested in clustering 4 color including one white background and hence k=4. Lets this be (66, 76, 45), (240, 230, 44), (44, 113, 110) and (5, 5, 200). This is entered in the separate cells as shown in the figure 14.

Cluster Center 1			Cluster Center 2			Cluster Center 3			Cluster Center 4		
R	G	B	R	G	B	R	G	B	R	G	B
66	76	45	240	230	230	44	113	110	5	5	200

Figure 14. Cluster Center's (Initially assumed)

The data points which has to be clustered here is the pixel values of the image (Apple). Now for convenience the pixel values obtained for the image is printed on to a single column excel file as shown in figure 16 (red, green, blue pixels values should be in different columns). The available data is written in the format to help "what if analysis".

TEST DATA			
SL NO	R	G	B
1	255	255	250
2	255	255	253
3	255	254	255
4	255	254	255
5	254	255	255
6	252	255	255
7	254	255	255
8	255	255	255
9	255	254	255
10	253	251	252
⋮	⋮	⋮	⋮
6895	255	255	255
6896	255	255	255
6897	255	255	255
6898	255	255	255
6899	255	255	255
6900	255	255	255

Figure 15. Test data

Now just as mentioned in k-means implementation description, we make an arrangement, where in the R, G and B values are acquired from the table in figure 15 by providing the index values. This arrangement is shown in figure 16.

INDEX OF CURRENT POINT	R	G	B
1	255	255	250

Figure 16. A requirement for what if analysis

Now as mentioned in normal K–means implementation we obtain the distance of the each data point (pixel value) from all the assumed centroid point. For each data point, find the minimum among the distances to all the four clusters and print it in the last column of the table in figure 17. The sum of these, minimum distances is found out and this is the objective function (value) which is to be minimized. Figure 17 illustrates the whole process of obtaining the Total distance to be minimized.

INDEX	Distance to CENTROID 1	Distance to CENTROID 2	Distance to CENTROID 3	Distance to CENTROID 4	CLASS	Minimum Distance
1	3.855842749	364.7653585	335.5164237	247.1443263	Cluster1	3.855842749
2	1.071822582	366.8152021	337.7469462	250.0180256	Cluster1	1.071822582
3	1.426063285	367.5032996	338.8809799	251.729982	Cluster1	1.426063285
4	1.426063285	367.5032996	338.8809799	251.729982	Cluster1	1.426063285
5	1.384192757	367.9557345	338.6846875	251.7444007	Cluster1	1.384192757
6	2.782442756	367.4969139	337.5790985	251.373779	Cluster1	2.782442756
7	1.384192757	367.9557345	338.6846875	251.7444007	Cluster1	1.384192757
8	1.415270483	368.1890045	339.2405525	251.935461	Cluster1	1.415270483
9	1.426063285	367.5032996	338.8809799	251.729982	Cluster1	1.426063285
10	4.203614173	362.9107399	334.4479302	247.867594	Cluster1	4.203614173
⋮	⋮	⋮	⋮	⋮	⋮	⋮
6895	1.415270483	368.1890045	339.2405525	251.935461	Cluster1	1.415270483
6896	1.415270483	368.1890045	339.2405525	251.935461	Cluster1	1.415270483
6897	1.415270483	368.1890045	339.2405525	251.935461	Cluster1	1.415270483
6898	1.415270483	368.1890045	339.2405525	251.935461	Cluster1	1.415270483
6899	1.415270483	368.1890045	339.2405525	251.935461	Cluster1	1.415270483
6900	1.415270483	368.1890045	339.2405525	251.935461	Cluster1	1.415270483
					Total	190328.174

Figure 17. Total distance to be minimized

Now we utilize "Solver" to find out an optimal value for R, G and B for each of the centroids. The optimum values of the centroid points found by minimizing the sum of distances to the nearest centroid points is shown in the below figure.

Cluster Center 1			Cluster Center 2			Cluster Center 3			Cluster Center 4		
R	G	B	R	G	B	R	G	B	R	G	B
254.4565	254.515333	253.786453	168.6397843	2.26609237	1.56237829	66.0825561	132.583061	1.21181	206.383	202.754	13.3841

Figure 18. Optimized centroid values

Due to use of "What-If Analysis", once the cluster centers (centroids) are changed the allocation of data point to the clusters also changes automatically. Now we plot the segmented image of "Apples" by assigning different user interested colors to the points

belonging to different clusters. This is also done with the help of a macro. The segmented image is obtained as shown in figure 19.

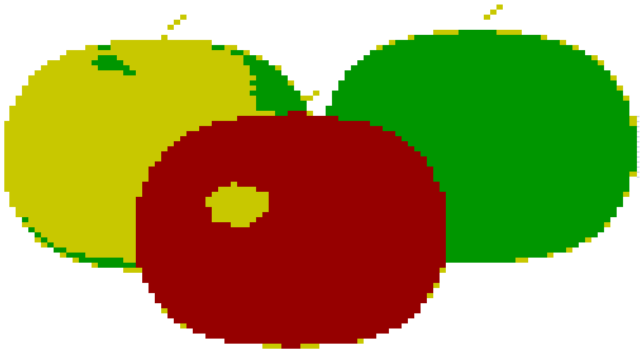

Figure 19. Clustered / Segmented Image

3. CONCLUSION

Data Clustering or grouping together of similar data comes useful in all real world data processing applications. This paper deals with the implementation of k-Means clustering in Microsoft Excel 2007 with the help of "What-If Analysis" and "Solver" add-in packages available by default. As an application of clustering, this paper explains, how image segmentation can be done in Microsoft Excel. The paper also explains on newly developed add-in which can read any image and obtain the pixel values for red, green and blue component of the image. Thus it is shown that the whole set of image processing operations such as reading, processing and printing of image can be done in excel. Above all this paper stressed on the potential of Microsoft Excel as a scientific learning tool.

4. REFERENCES

[1] Ali, R., Ghali, U., Saeed, A. "Data Clustering and Its Applications". Available at the web address : http://members.tripod.com/asim_saeed/paper.htm

[2] Gao, J., Goodmen, Joshua T., Miao,J. "The Use of Clustering Techniques For Language Modeling". International Jouranal for Computational Linguistics and Chinese Language Processing. Vol. 6, No. 1, pp 27-60.

[3] Gunnerson, E. and Wienholt, N. (2005), A Programmer's Introduction to C # 2.0, Third Edition, Published by Apress. ISBN (pbk) : 1-59059-501-7

[4] Implementation of clustering in Excel and Excel Add-in to load Images. Available at the URL : http://cen.amritafoss.org/downloads/Books/DataMining/Worksheet/ .

[5] MacDonald, M. (2006), Excel 2007 : The Missing Manual.Published by Pogue Press, O'Reilly. ISBN:978-0-596-52759-4.

[6] Ragsdale, C. T. and Zobel, C. W. (2010), A Simple Approach to Implementing and Training Neural Networks in Excel. Decision Sciences Journal of Innovative Education, 8: 143–149. doi: 10.1111/j.1540-4609.2009.00249.x

[7] Soman, K.P., Loganathan, R., Ajay, V. Machine learning With SVM and Other Kernel Methods. Published by PHI Learning Private Limited. ISBN:978-81-203-3435-9

[8] Tang, H. (2008), A Simple Approach of Data Mining in Excel.IEEE Fourth International Conference Wireless Communications, Networking and Mobile Computing, doi : 10.1109/WiCom.2008.2679