Paper CI-08

A Brief Survey of Clustering and Segmentation for Customer Intelligence

George J. Hurley, The Hershey Company, Hershey, PA

Abstract

Clustering and segmentation is important in nearly all industries as it allows businesses to identify and categorize customers. Using this approach a business can market and develop products for a targeted segment(s). The methods of these targeted objectives are dependent on business need, industry and type; that is, one industry may use a direct mail campaign, while another uses an ad campaign to one segment, and develops a new product for another. Regardless of the business implementation, the underlying methodology of segmentation is fairly similar. In this paper, implementation and basic technical aspects of hierarchical clustering, k-means clustering, normal mixtures, and self-organizing maps will be discussed. These methods are all available in JMP® Pro. In addition to comparing these methodologies, the paper will also discuss some common business uses of clustering.

Clustering

Simply put, clustering exists in order to separate things into groups. The idea of clustering in a basic form is perhaps older than history, when some ancient person decided that there are similarities and differences in groups of objects.

Consider a group of seven coins on a table: three quarters, three pennies, and a dime. Consider also the author's four children (aged 3, 6, 13, and 13).

The three year old can place the seven coins into three groups and can tell you why… "three are large and four are small". Of those four that are small, he'd continue, "three are copper colored" (although I suspect he might say brown), and "one is silver". This can be done by the child without them having any knowledge of what their monetary value is or of the words on them.

The six year old, then, upon review, may identify that two of the quarters are traditional quarters with George Washington on one side and an eagle on the other, while the third is a state quarter, having perhaps an airplane and an astronaut on the "tails" side. Hence she may separate the quarters into four clusters, claiming an improvement.

The first teenager may then note that the two eagle coins are from several different years, and that one of the pennies is also of a different year of than the other two. Hence, he may break the pile into six clusters, separating the quarter, and then the penny.

Finally, the last teenager may note the condition of wear is different on the two pennies of the same year, separating them and forming seven clusters. At this point, all four children will have come to different results, and would likely begin arguing over whose clustering was correct, to which the author would need to intervene and let them know that no answer is wrong.

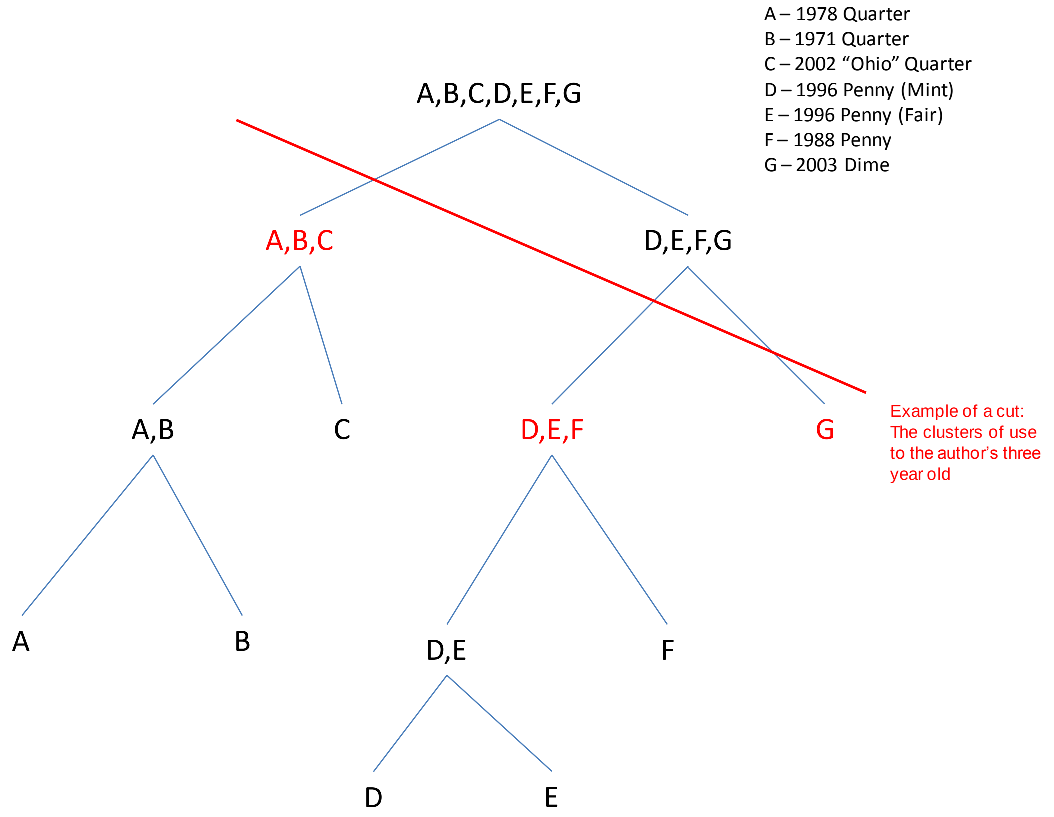
There are three important lessons from this anecdote.

1) For any given problem, there are multiple solutions. Some are more useful than others, depending on the situation.
2) Clustering can use continuous/ordinal (size of coin/year of production) or categorical (color of coin/condition of coin) data. However, in mathematical practice, categorical data is typically recoded as multiple binary variables or ordinal, as appropriate.
3) If data is provided any set of objects can be clustered. You do not need the physical objects present (the reader should have followed the anecdote without having to take out coins). This is why clustering is in the realm of analytical sciences.

The example given above is an example of divisive hierarchical clustering.

Hierarchical clustering is typically presented visually in a dendogram, which shows graphically the splits (or agglomerations in the case of agglomerative hierarchical clustering) that lead to the clusters. Figure 1 illustrates a dendrogram for the example above. It can be noted that the dendrogram can be "cut" anywhere of use to provide clusters.

Figure 1: A Dendrogram of the Coin Clustering



Hierarchical Clustering

Hierarchical Clustering refers to clustering methods that attempt to break data into a hierarchy of clusters. Generally this can be done in two directions.

First, one can consider divisive hierarchical clustering, where one begins with all items in one giant cluster and divides (splits) them into subsequent clusters until some criteria is met. Generally, a distance metric is chosen and the split that maximizes the distance between the two newly formed clusters is the split that is chosen.

Alternatively, one can use agglomerative clustering, where each item begins in its own cluster and clusters are subsequently joined until a criterion is met. Generally, in the case of agglomerative clustering, a distance metric is chosen such that the two clusters which are most near each other initially are chosen for the join.

For hierarchical clustering, any valid measure of distance can be used[1] as a distance metric. Generally speaking, there are two concepts at play in calculating a distance. First, one must determine the linkage. That is what method will be used to describe the relationship of a **group of objects** to another **group of objects**. That is, one may say the distance between one group and another group is the distance between their closest members, which is called single linkage. One alternative method, called complete linkage may be to say it is the distance between their furthest members. There are various methods for linkage available. Once linkage is determined, the metric used to calculate distance must be considered. For example, Euclidean distance is and squared Euclidean distance is . Various linkage and metrics for distance are presented at http://en.wikipedia.org/wiki/Hierarchical_clustering.

JMP performs agglomerative hierarchical clustering. It offers several different methods for calculating distance in hierarchical clustering. Specifically, JMP offers Average Linkage, Centroid Method, Ward's Minimum Variance, Single Linkage, Complete Linkage, and Fast Ward. The technical specifications of these are available in the help file for JMP. Ward's Minimum Variance is the default method chosen by JMP. In this author's opinion, this is a good option, although it is sensitive to outliers. Taking directly from the JMP help file[2],

In Ward's minimum variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give the proportions of variance (squared semipartial correlations).
Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the assumptions of multivariate normal mixtures, spherical covariance matrices, and equal sampling probabilities.

Ward's method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with approximately the same number of observations. It is also very sensitive to outliers. See Milligan (1980)

Based on the same documentation, the centroid method is more robust to outliers[2]. However, by default JMP also selects "Standardize Data" as an option. Standardizing data before clustering will reduce the effects of outliers and different variable scales.

Example – Hierarchical Clustering

To perform this clustering method, the famous Fisher's Iris dataset will be used. This data set contains three species of irises and measurements of their sepal length, sepal width, petal length, and petal width. The idea is to cluster the flowers into their correct species by these measurements. This is available in JMP under Help Sample Data.

Once opened, select Analyze… Multivariate Methods… Cluster to bring up the clustering window. The four petal lengths can then be chosen for Y. Note the default Option of "Hierarchical" is selected along with "Ward" and "Standardize Data". Once ok is selected, the output in Figure 2 is produced. It is immediately clear from the dendrogram shading that there are three main groups, but there is some mixing, and if the clusters are output by clicking on the red triangle and selecting "Save Clusters", then there were actually 13 generated.

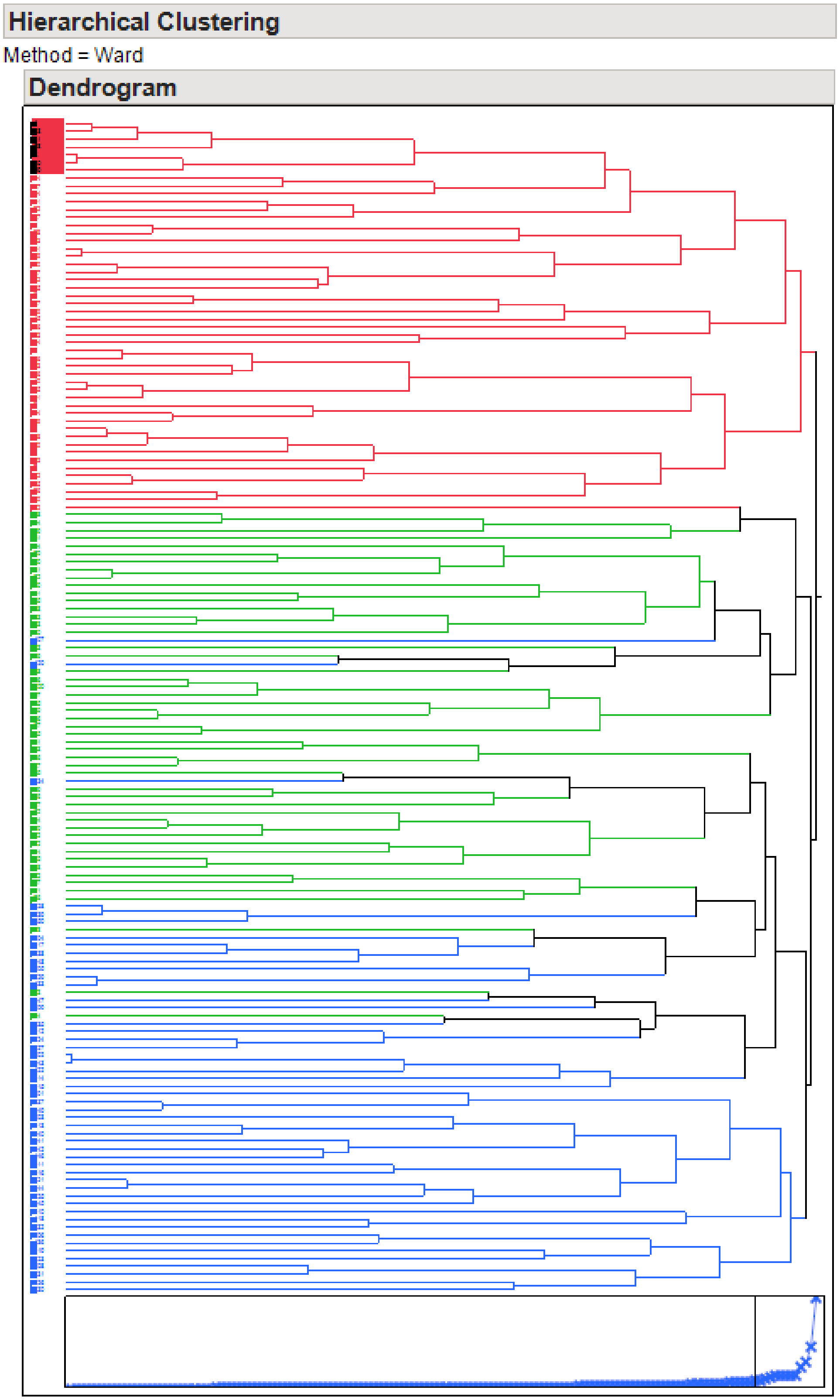
Since it is known that 3 clusters are desirable, rather than allowing JMP to use its internal mechanism for determining cluster number, under the red triangle, check "Color Clusters" and "Mark Clusters" and then set "Number of Clusters" to 3 (or any other number if you have more or less). The dendrogram will update to view the clusters cleanly, while the clusters become marked in the data. Further, when "Save Clusters" is selected, all observations will be assigned to one of the three clusters. Figure 3 illustrates that the clustering method does well for Virginica and Setosa, but has issues differentiating Versicolor from Virginica, that is it identifies a sizable number of the Versicolors as Virginicas. It should be noted that the plot symbol in Figure 3 originates from marking the data with the clusters using "Mark Clusters" as described. Consider Figure 4, which plots the four cluster solution as symbols on the plot from Figure 3. What is seen is that the fourth cluster is clearly identified Virginicas, while the third cluster (blue) is still somewhat mixed.

K-Means Clustering

While hierarchical clustering is a useful methodology, it is can be regarded as simplistic by some researchers, this author included. This is in part due to the fact that many newer methodologies perform better. While K-Means is not a new methodology anymore, it is still a very useful clustering methodology.

Rather than relying on recursive splitting or agglomeration of data points, K-Means initially chooses a number of points equal to the desired cluster number with the domain of the data. These are called "cluster seeds" and are typically chosen at random. Observations are then assigned to the cluster seed they are located closest to. The seeds are then moved to the centroid of the set of observations that were assigned to them (and are now called cluster means). Observations are then reassigned to the cluster mean they are closest to. This process is repeated until a threshold is met and all observations are assigned to a cluster.

Figure 2: A Dendrogram of Irises from JMP



Hierarchical Clustering

Method = Ward

Dendrogram

Figure 3: 3 Clusters and Actual Species

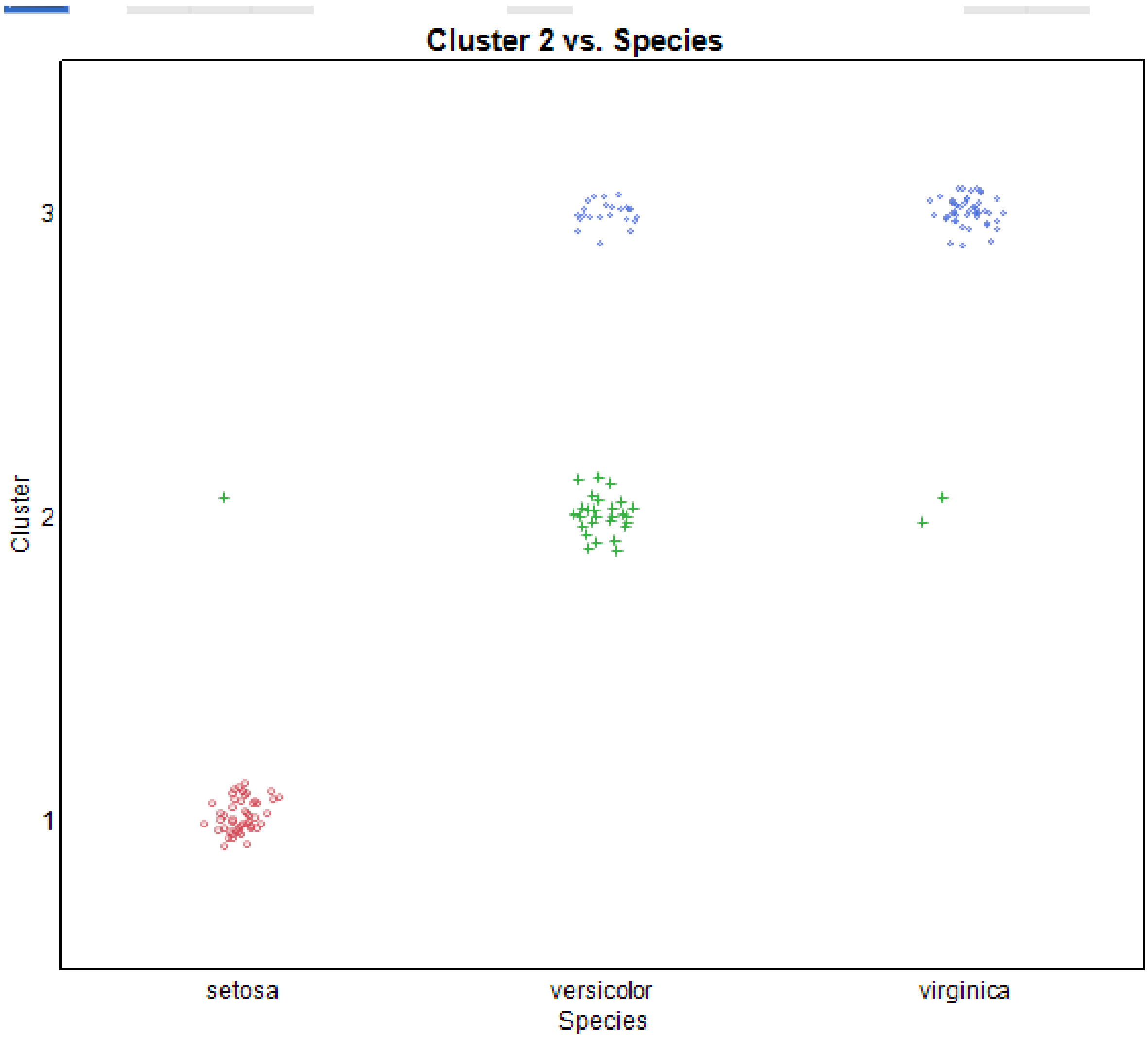
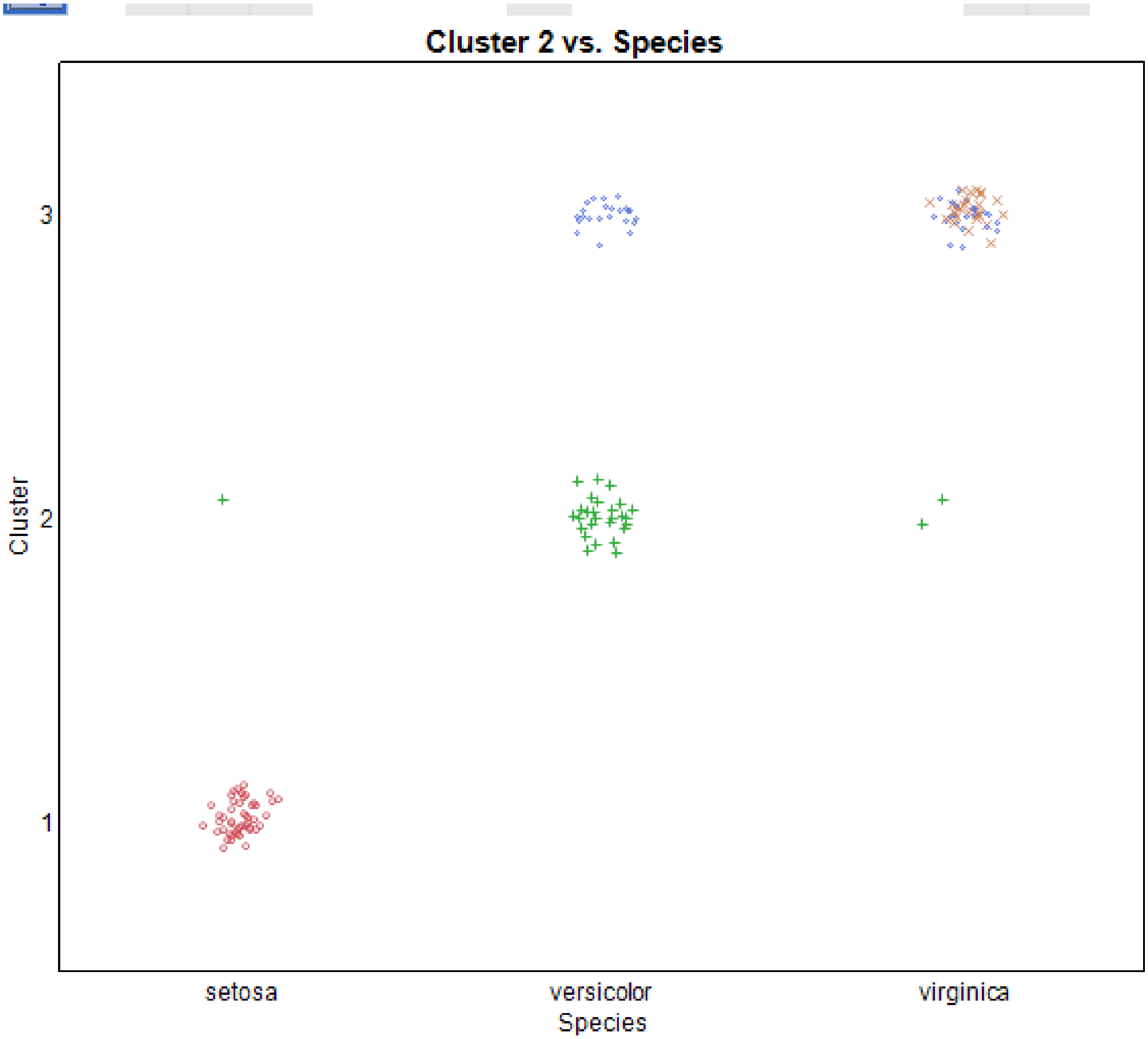


Figure 4: 4 Clusters and Actual Species



K-Means Clustering Example

Again, the Fisher Iris data set will be considered. In JMP, the clustering window will once again be opened, but rather that "Hierarchical"; "K-Means" will be chosen. This brings up a window where again, the four petal lengths will be chosen for Y. The window allows for two options, "Columns Scaled Individually", which is by default checked, and "Johnson Transform", which is by default unchecked. It is good practice to leave "Columns Scaled Individually" checked, as it will address issue of differing variable scales. Johnson Transform is used to address the issue of outliers. Here, Johnson Transform is left unchecked.

Once "OK" is clicked, the K-Means Cluster window (also known as the Iterative Clustering Panel) is then shown. This gives several options. The first is "Declutter". Declutter can be used to identify outliers. It will not be discussed here, but is presented in the JMP 10.0.0 help document.

"Method" allows the choice of various clustering methods. Here, since K-Means is being chosen, it will be left to its default, "K-Means Clustering". Next, there are two boxes, "Number of Clusters…" and "Optional range of clusters". Often, in practice, it is unknown how many clusters there are. For example, consider consumers purchasing automobiles. There may be three clusters of consumers, one preferring trucks, one preferring cars, and one preferring vans. However, perhaps there are four, where the car group is split between sports cars and family cars. It is for such exploratory analysis that this option proves very useful. In this case, "Optional range of clusters" is left blank, as it is known there are three types of flowers.

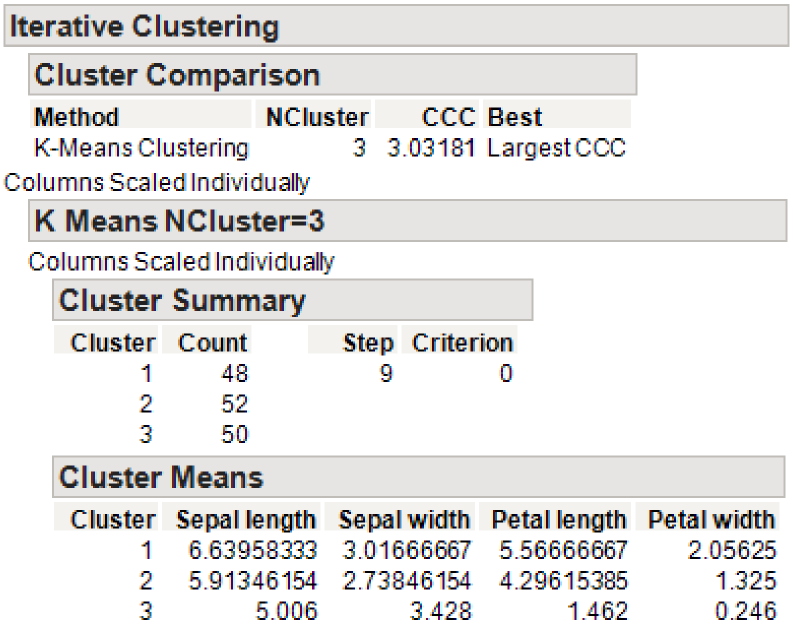
Finally, near the bottom, there are three checkboxes, "Single Step", "Use within-cluster std deviations", and "Shift distances using sampling rates". "Single Step" allows the user to step through the clustering process. "Use within-cluster std deviations" allows distances to be scaled based on the standard deviation of each cluster. If unchecked, distances are scaled by the overall standard deviation. Using this option will allow tight clusters to remain tight. "Shift distances using sampling rates" assumes unequal sized clusters and gives preference to assigning points to larger clusters. Here all three will remain unchecked. Once "Go" is selected, a report is given showing the number of observations assigned to each cluster and what the variable means are for each cluster. Figure 5 illustrates these results and shows that they are promising. Selecting "Save Clusters" from the red triangle near "K Means NCluster=3" will save the cluster assignment to the data table. From the same red triangle, "Save Colors to Table" can be selected to assign cluster colors to the data table (note it is useful to clear row states first). Figure 6 illustrates that the clusters still are much improved, with about six missed.

It is here that it is appropriate to note for those new to this dataset, that this well known dataset has a well known difficulty; the Virginica and Versicolor irises are difficult to differentiate, as their sizes are much more similar than either compared to the Setosa. The size differences are easily seen in Figure 5.

Normal Mixtures Clustering / Robust Normal Mixtures Clustering

Normal Mixtures clustering is a method that is designed to calculate the probability that each row is in a cluster, rather than "grouping rows"[2]. It proceeds by predicting the "proportion of responses expected

Figure 5: K-Means Clustering Report



Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	3	3.03181	Largest CCC

Columns Scaled Individually

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	48
2	52
3	50

Step	Criterion
9	0

Cluster Means

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	6.63958333	3.01666667	5.56666667	2.05625
2	5.91346154	2.73846154	4.29615385	1.325
3	5.006	3.428	1.462	0.246

within each cluster."[2] Further, it assumes "that the joint probability distribution of the measurement columns can be approximated using a mixture of multivariate normal distributions, which represent different clusters. The distributions have mean vectors and covariance matrices for each cluster."[2]

What is ultimately generated is a probability that each observation belongs to a cluster. This is an especially useful method when clusters may overlap. This method is sensitive to outliers, but a variation of the method, known as Robust Normal Mixtures Clustering, addresses these issues.

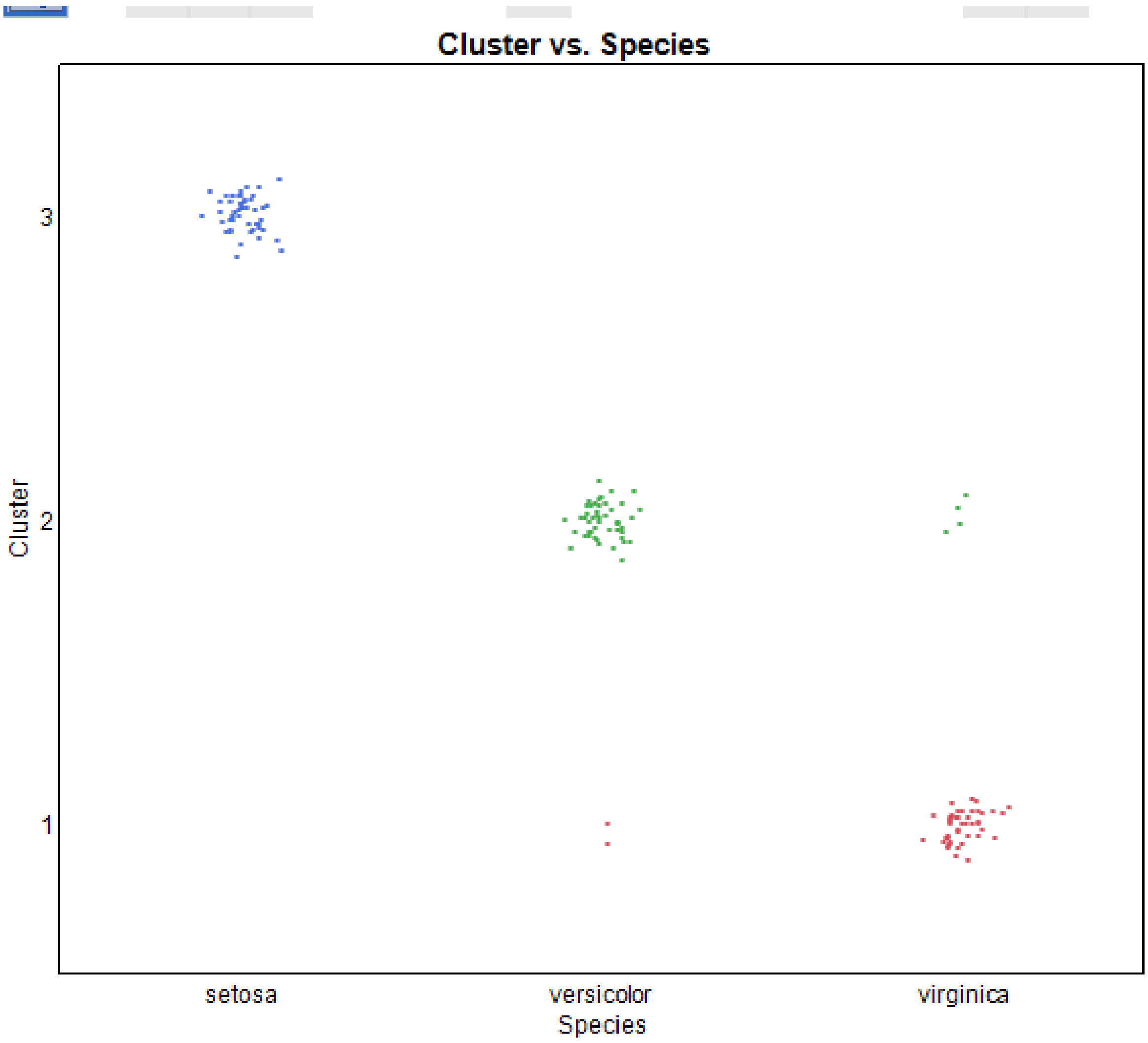
Normal Mixtures Clustering / Robust Normal Mixtures Clustering Example

Again, the Fisher Iris dataset is considered. To perform Normal Mixtures Clustering in JMP, repeat the same procedure as K-Means, but on the K-Means Cluster window, select "Normal Mixtures" or "Robust Normal Mixtures" under method.

First, Normal Mixtures will be considered. There are five options. Diagonal Variance is used to force off diagonal covariance matrix elements to zero. This forces the multivariate normal distributions fit to have no correlations between variables. It is useful to "avoid getting a singular covariance matrix, when there are fewer observations than columns."[2]

Outlier Cluster specifies that a cluster be created to catch outliers that do not fall in the other clusters.

Figure 6: K-Means Clusters and Actual Species



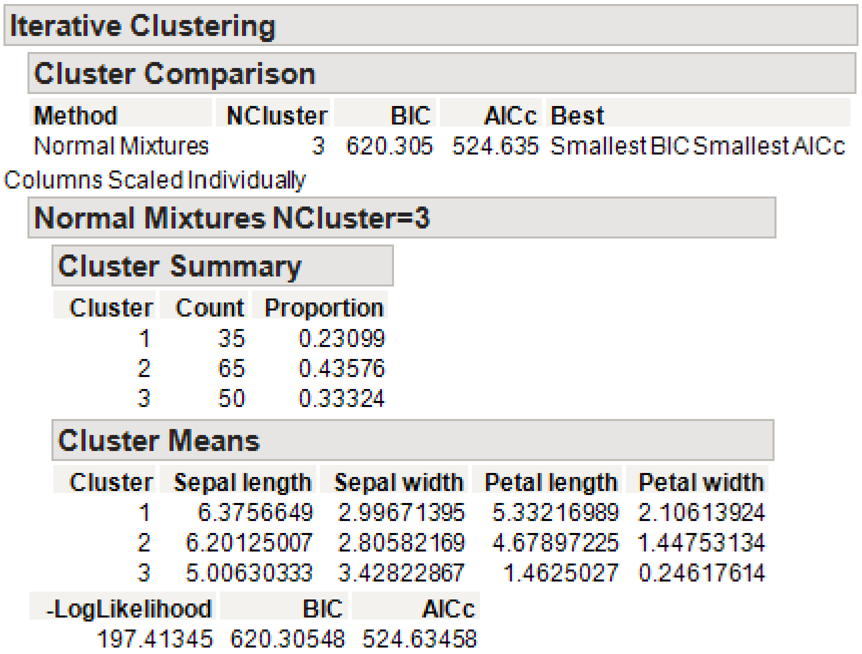
Outlier Cluster specifies that a cluster be created to catch outliers that do not fall in the other clusters. This is a useful option, especially in exploratory analyses.

Tours is how any times the estimation process is restarted with different initial values. A high number of tours will be computationally expensive, but will reduce the likelihood of local solutions.

Maximum Iterations "is the maximum number of iterations of the convergence stage of the EM algorithm."[2] Again, higher values are more computationally expensive, but may produce more stable results.

Convergence Criteria "is the difference in the likelihood at which the EM iterations stop."[2] This author has found that it is generally reasonable to use the default here.

Figure 7: Normal Mixture Clustering Report – Default Options



Iterative Clustering

Cluster Comparison

Method	NCluster	BIC	AICc	Best
Normal Mixtures	3	620.305	524.635	Smallest BIC Smallest AICc

Columns Scaled Individually

Normal Mixtures NCluster=3

Cluster Summary

Cluster	Count	Proportion
1	35	0.23099
2	65	0.43576
3	50	0.33324

Cluster Means

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	6.3756649	2.99671395	5.33216989	2.10613924
2	6.20125007	2.80582169	4.67897225	1.44753134
3	5.00630333	3.42822867	1.4625027	0.24617614

-LogLikelihood	BIC	AICc
197.41345	620.30548	524.63458

For this example, all defaults will be used. Once the process is run, again Clusters are saved and Colors are stored to the Table. Figure 7 and Figure 8 display the results. It can be seen that the standard Normal Mixtures with the defaults did not improve the output.

However, if the process is reran with number of tours set to 30000 and number of iterations set to 3000, Figure 9 and Figure 10 can be generated, showing the best results seen to this point, with only two observations misallocated.

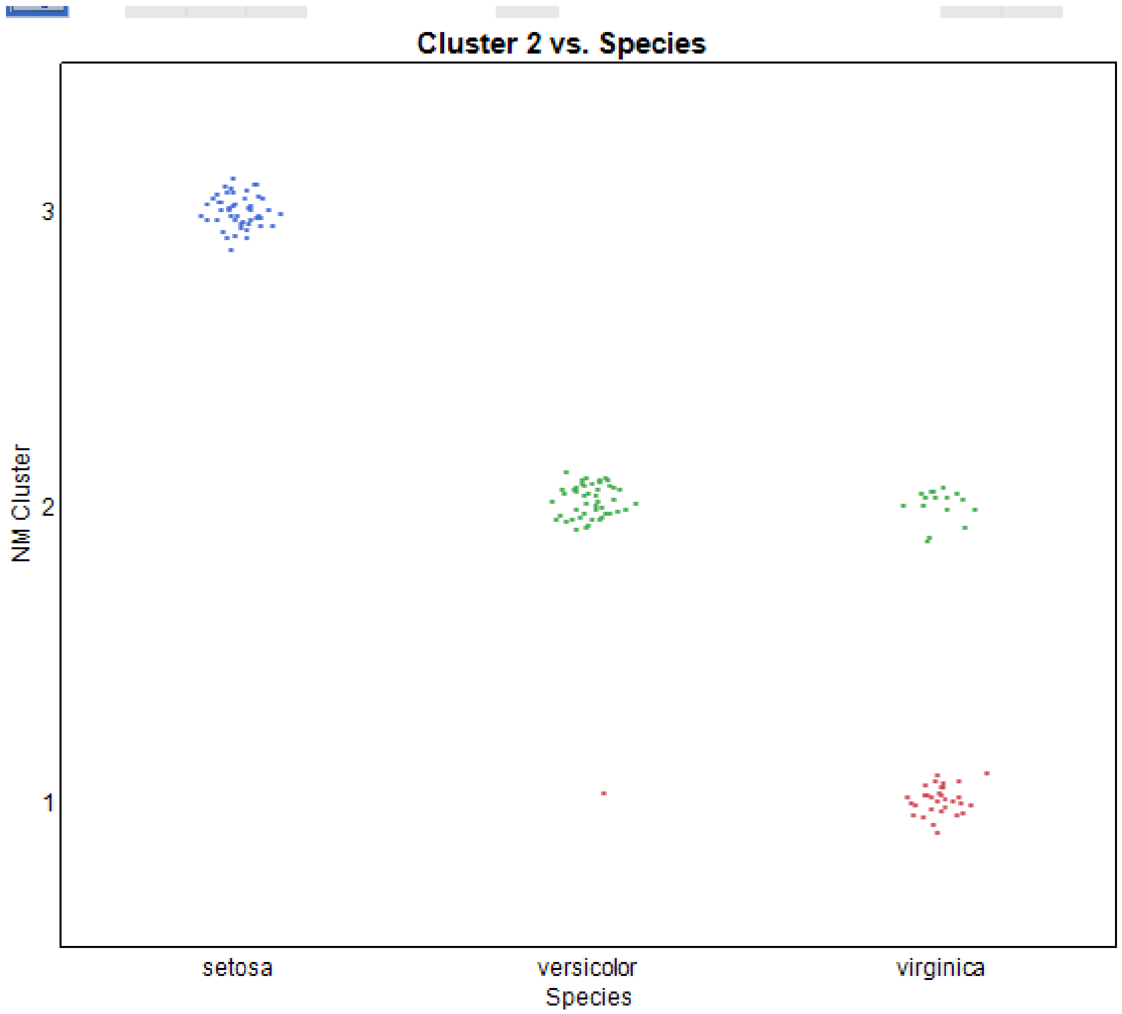
This illustrates the importance of having sufficient tours to find a globally optimal solution.

Robust Normal Mixtures will now be considered. Again, on the K-Means Cluster window, select “Robust Normal Mixtures” under method. There are five options. Diagonal Variance has the same meaning as under Normal Mixtures.

Robust Normal Mixtures down weights outliers. Roughly speaking, choosing a number for the Huber coverage that is close to 1 protects only against the most extreme outliers, while numbers closer to 0 suggest a larger proportion of the data be down weighted. Too small a choice may lead to non-outliers being considered outliers, while too large a number may exclude some outliers. In most exploratory analyses, it is appropriate to choose a relatively high value. The default is 0.9 and this author has found it quite acceptable in practice.

Complete Tours is how any times the estimation process is restarted with different initial values. A high number of tours will be computationally expensive, but will reduce the likelihood of local solutions.

Figure 8: Normal Mixture and Actual Species – Default Options

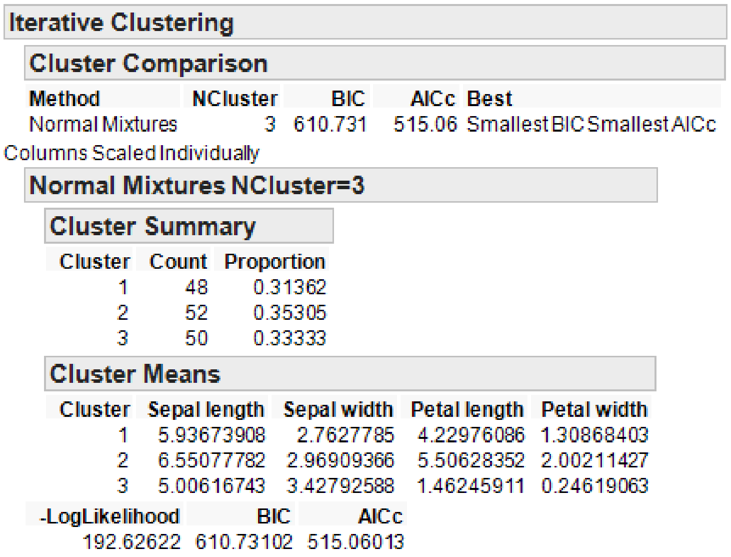


“Initial guesses is the number of random starts within each tour. Random starting values for the parameters are used for each new start.”[2]

“Max Iterations is the maximum number of iterations during the convergence stage. The convergence stage starts after all tours are complete. It begins at the optimal result out of all the starts and tours, and from there converges to a final solution.”[2]

Like before, we generate the same report and graphs based on the Fisher data, and utilize the defaults for Robust Normal Mixtures (which includes 10 complete tours). These are seen in Figure 11 and Figure 12. It is noted that the results are roughly on par with K-Means, but not as good as the Normal Mixture with 30,000 tours. This again emphasizes the value of taking many tours to find a global solution. However, it should be noted that increasing tours is computationally expensive and it is advisable to run many tours on a powerful machine.

Figure 9: Normal Mixture Clustering Report – 30,000 Tours



Iterative Clustering

Cluster Comparison

Method	NCluster	BIC	AICc	Best
Normal Mixtures	3	610.731	515.06	Smallest BIC Smallest AICc

Columns Scaled Individually

Normal Mixtures NCluster=3

Cluster Summary

Cluster	Count	Proportion
1	48	0.31362
2	52	0.35305
3	50	0.33333

Cluster Means

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	5.93673908	2.7627785	4.22976086	1.30868403
2	6.55077782	2.96909366	5.50628352	2.00211427
3	5.00616743	3.42792588	1.46245911	0.24619063

-LogLikelihood	BIC	AICc
192.62622	610.73102	515.06013

Figure 10: Normal Mixture Clustering 30,000 Tours and Actual Species

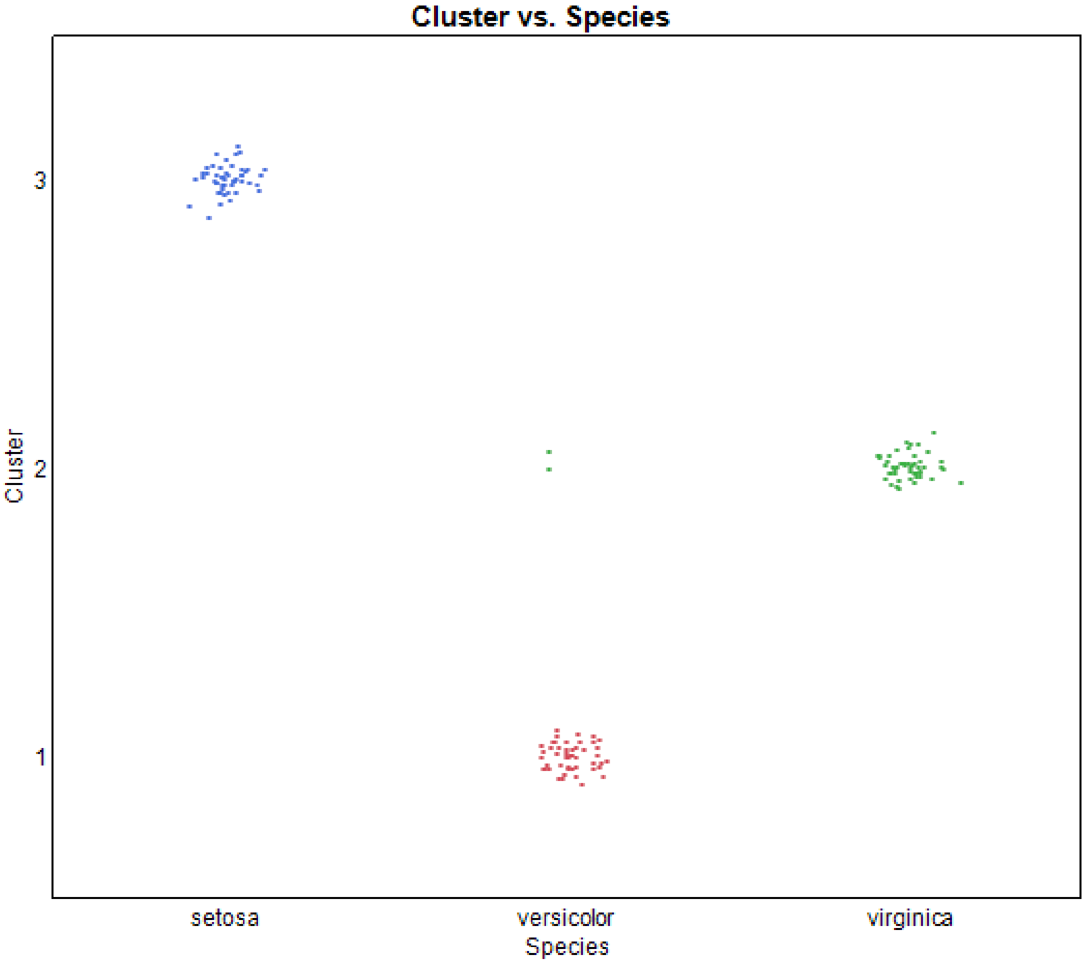
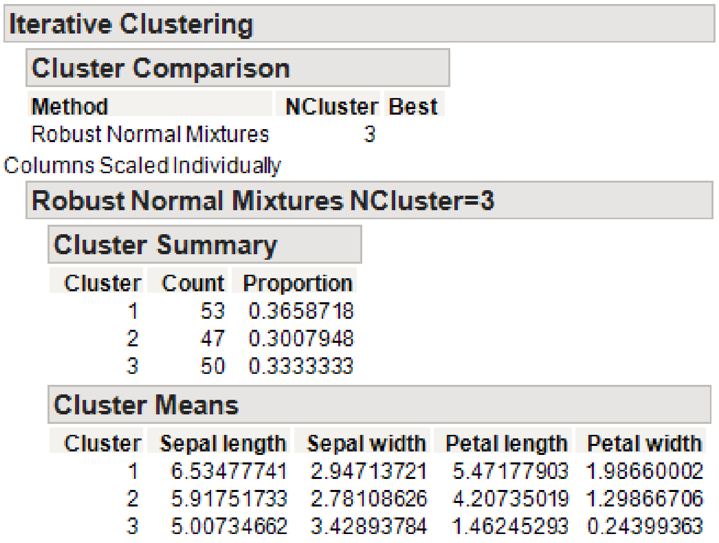


Figure 11: Robust Normal Mixture Clustering Report



Iterative Clustering

Cluster Comparison

Method	NCluster	Best
Robust Normal Mixtures	3	

Columns Scaled Individually

Robust Normal Mixtures NCluster=3

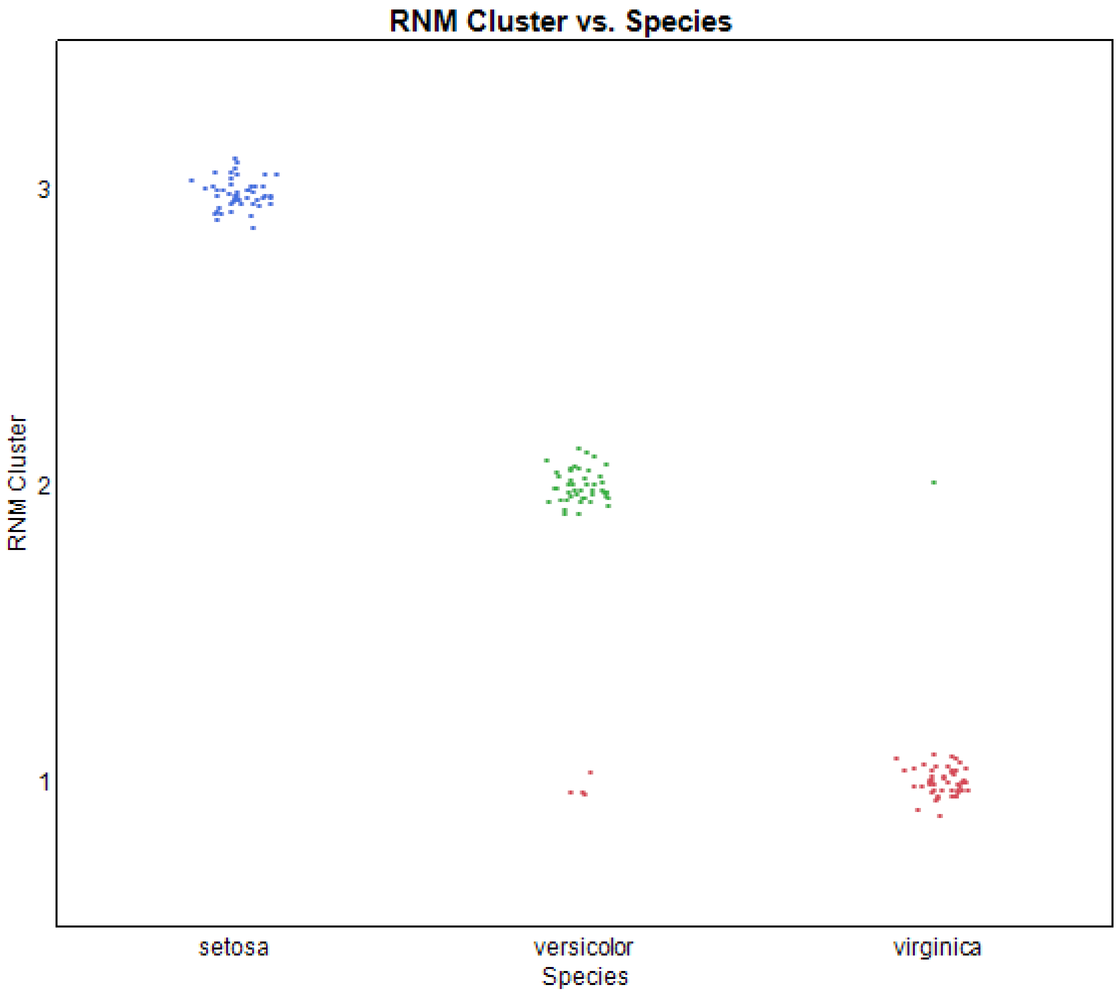
Cluster Summary

Cluster	Count	Proportion
1	53	0.3658718
2	47	0.3007948
3	50	0.3333333

Cluster Means

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	6.53477741	2.94713721	5.47177903	1.98660002
2	5.91751733	2.78108626	4.20735019	1.29866706
3	5.00734662	3.42893784	1.46245293	0.24399363

Figure 12: Robust Normal Mixture Clustering and Actual Species



Self Organizing Map

The final method that JMP offers for clustering is the Self Organizing Map. The Self Organizing Map is a method that was designed to give low dimensional views to high dimensional data. Some implementations of Self Organizing Maps use Artificial Neural Networks to create the map.[3] In JMP, the implementation of the Self Organizing Map is not through ANNs, but rather through what is termed a "much more straightforward way as a simple variation on k-means clustering. In the SOM literature, this would be called a *batch algorithm* using a *locally weighted linear smoother*."

As implemented in JMP, they essentially create clusters that have spatial properties in a grid. That is, if the clusters are imagined in a grid, those clusters near each other on the grid are also near each other in multivariate space.

The Self Organizing Map is useful because it allows the researcher to understand the inter-cluster distances in a meaningful spatial way. For example, if there are many clusters of customers to be targeted via direct mail, it may be useful to understand which ones are spatially nearer each other in 2-space to combine, if there are limited funds to develop targeted messages to mail to each segment. Of course, the goal of combining customers can be met by all methods above in different ways.

Self Organizing Map Example

Here, again, the Fisher data is used. Self Organizing Map (SOM) is selected from the K-Means Cluster window. There are four options available.

Single Step allows the user to step through the clustering process.

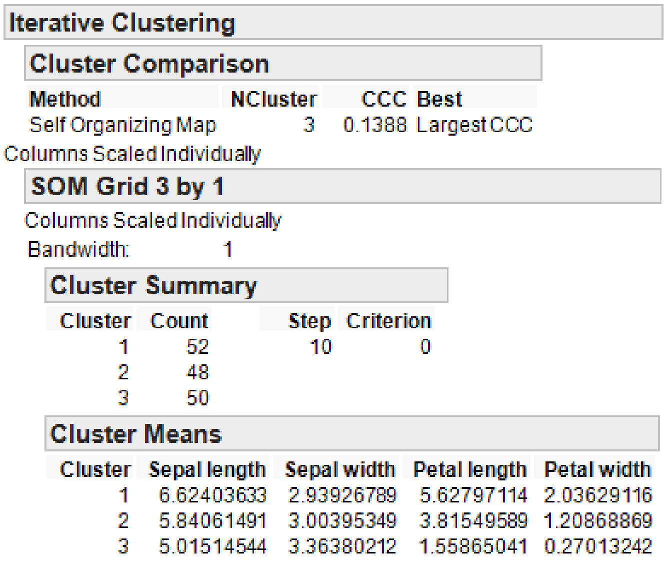
N Rows and N Columns allow the user to layout the grid. There will be N Rows*N Columns clusters produced.

"Bandwidth determines the effect of neighboring clusters for predicting centroids. A higher bandwidth results in a more detailed fitting of the data."[2]

In this example, since 3 clusters are desirable, N Rows is set to 1 and N Columns is set to 3. Here, the default bandwidth is changed, as a more detailed fitting of the data is desirable, given the similarity of the Virginica and Versicolor species. The bandwidth is therefore set to 1. Go is then selected.

Figure 13 and Figure 14 illustrate that a reasonable fit is seen from using the SOM. Also note that cluster 1 and 3 are the furthest apart from each other. Since this grid is really 1-dimensional, this is apparent. For 2-dimensional grids, it is useful to layout the clusters in the grid as it was specified with N Rows and N Columns to examine the relationships revealed.

Figure 13: Robust Normal Mixture Clustering Report



Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
Self Organizing Map	3	0.1388	Largest CCC

Columns Scaled Individually

SOM Grid 3 by 1

Columns Scaled Individually

Bandwidth: 1

Cluster Summary

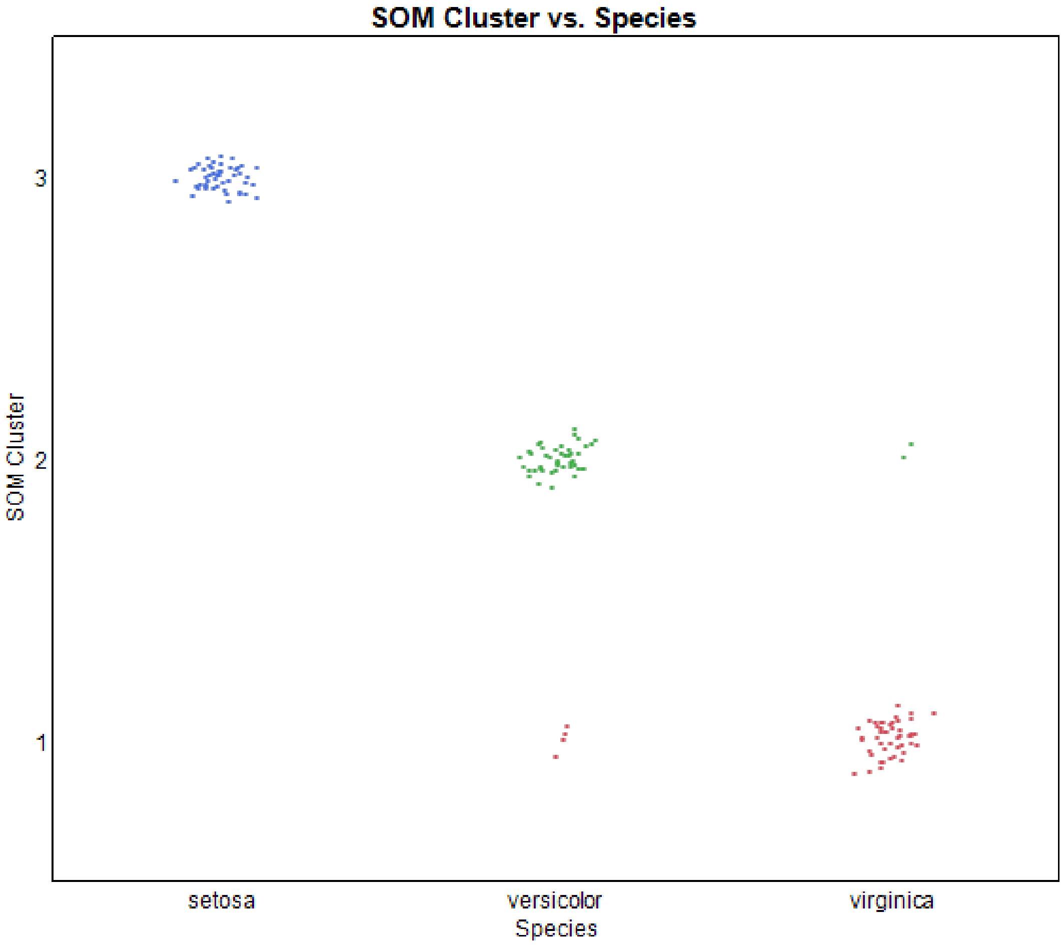
Cluster	Count
1	52
2	48
3	50

Step	Criterion
10	0

Cluster Means

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	6.62403633	2.93926789	5.62797114	2.03629116
2	5.84061491	3.00395349	3.81549589	1.20868869
3	5.01514544	3.36380212	1.55865041	0.27013242

Figure 14: Robust Normal Mixture Clustering and Actual Species



Discussion / Ideas for Usage

It should be noted that each clustering method has goodness of fit statistics and other related metrics. However, it is often the case that the most desirable outcomes are predictive power of an existing set, possibly paired with a validation set. In the case discussed ad infinitum, Fisher could have potentially generated a second validation sample that the 30,000 tour Normal Mixture Cluster formula could have been applied to. If this formula showed good predictive properties for the validation set, it would logically follow that it would be useful in predicting what type of iris a particular sample was. This is more akin to a typical business situation, where there is a desired endpoint. For example, replace "iris" with "customer" and then act on the predicted samples (i.e. newly segmented prospective customers) with a targeted message, and a typical business use arises.

In one example, it may be useful for a bank issuing a direct mail campaign to be able to differentiate their potential clients into segments that will each get messaged differently from marketing.

In this case, it may be useful to generate varying numbers of segments using the various methods to determine which ones have desirable properties in describing customer's attributes in a business sense from an existing client database. For example, in this case, perhaps they fall into high net worth clients, high fee generators, and low revenue customers.

Once it is determined who is in these categories of the existing clients, either a predictive model can be generated to predict the segment based on publically available data for the potential clients, or, alternatively, the segments can be recreated using these segmentation methods from the publically available data, perhaps in a test and validation dataset setting. The segment formulas can then be saved and applied to the potential client base.

Another example may be in a company that produces cars. This company may want to segment its customers using one of these methods. The segments may reveal a business need. For example, perhaps one segment hasn't bought a new car from the company in over ten years. Perhaps this group also in high numbers bought a particular model that was highly affordable, but not known for reliability. It may be useful for the company to develop a new vehicle that is affordable, yet also reliable, and then target the lost customers with ads touting the vehicles reliability and price.

These are obviously contrived examples, but the idea behind segmentation is that it is general used to differentiate groups of things, which in a business setting are typically customers or groups of customers. It is also often used to project segments onto prospective customers in order to target them with specific messaging, generally with a cluster formula or a predictive model generated from a known databases segmentation. Finally, it can be used in new product development, as described above.

References

[1] http://en.wikipedia.org/wiki/Hierarchical_clustering
[2] JMP 10.0.0 Help
[3] http://en.wikipedia.org/wiki/Self_organizing_map

Contact Info

Please contact the author with any questions.

George J. Hurley
The Hershey Company
19 E Chocolate Ave.
Hershey, PA 17033
ghurley@hersheys.com

Trademark Citation