Ball State Undergraduate Mathematics Exchange
http://www.bsu.edu/libraries/beneficencepress/mathexchange
Vol. 12, No. 1 (Fall 2018)
Pages 35 – 50

Exploration of Common Clustering Methods and the Behavior of Certain Agreement Indices

Y. Huang and A. Flynt

Y. Huang graduated with a degree in mathematics from Bucknell University. He then attended Columbia University for graduate program in actuarial science. He now works as an actuarial analyst in the insurance industry, and is studying for the Fellowship of Society of Actuaries.

A. Flynt is an Assistant Professor in the Department of Mathematics at Bucknell University, teaching statistics and data science courses as well as a sports analytics course for incoming first years. Her research is primarily in statistical clustering with applications in medicine and social justice.

Abstract Statistical clustering is an exploratory method for finding groups of unlabeled observations in potentially high dimensional space, where each group contains observations that are similar to each other in some meaningful way. There are several methods of clustering, with the most common including hierarchical clustering, k-means clustering and model-based clustering. Agreement indices are quantitative metrics that compare two partitions or groupings of data. In this paper, we introduce three clustering methods and compare their results using different agreement indices, after being applied to Fisher's iris data, a classic clustering benchmark data set.

1 Introduction

The need to summarize data in an effective and efficient way is increasingly important because of the vast amount of "big data" now available in most disciplines, including the humanities, social and behavioral sciences, health and environmental sciences, and, of course, the natural sciences and engineering. Statistical clustering is an exploratory method which allows data to be summarized meaningfully by a relatively small number of clusters of objects or observations. Data in each cluster resemble each other and differ from data in

other clusters according to some criteria. If the data can validly be summarized by a small number of clusters, then the cluster labels may provide a very concise description of patterns of similarities and differences in the data.

A classification scheme may simply represent a convenient method for organizing a large data set so that it can be understood more easily and information can be retrieved more efficiently. A variety of clustering methods have been developed to accomplish this goal. In this paper, we discuss and compare hierarchical clustering, k-means clustering, and model-based clustering. These three clustering methods are the most commonly used due to their prominent advantages in data mining, but since they are different methods, they will often produce different clustering solutions on the same dataset. We begin the paper by introducing each method and their advantages and disadvantages. We then illustrate the clustering solutions produced by these methods for the benchmarking data set 'iris'. Finally, in order to compare these solutions, we utilize (and compare) three of the most commonly used agreement indices for comparisons of partitions of data. These include the adjusted Rand index, the Fowlkes-Mallows index, and the Jaccard index. Using these indices and a known species classification for the iris, we determine the optimal clustering solutions for the iris data.

2 Fisher's Iris Data

Fisher's famous iris data (Becker et al. [1988]) is a classic benchmarking dataset used in machine learning. The data is made up of sepal and petal measurements for 150 different iris flowers, 50 from each of three species: iris setosa, versicolor, and virginica. Specifically, we will try to cluster the iris into their respective species (true classification) using the sepal length and width and the petal length and width (all measured in centimeters) of each flower.

Figure 1, shows two pairsplots illustrating the bivariate relationships between the four quantitative variables, without and with the consideration of the species labels (colors). For all pairs of variables, two distinct groupings can be identified. Recall, however, that there are 3 species that we want to recover in the clustering. In the right pairsplot, the black, red and green observations represent the observations in the species setosa, versicolor, and virginica, respectively. The species setosa is well separated from the other species and should be easily distinguished as a cluster. The species versicolor and virginica have overlapping boundaries. Although in some subplots, these species appear to form a large mixed cluster, the observations of versicolor and virginica can still potentially be well discriminated. It can be inferred that the data intrinsically form 3 clusters, each representing the data of a species. This species classification will be used as the true clustering labels, and serve as our comparison in the evaluation of the performance of different clustering methods.

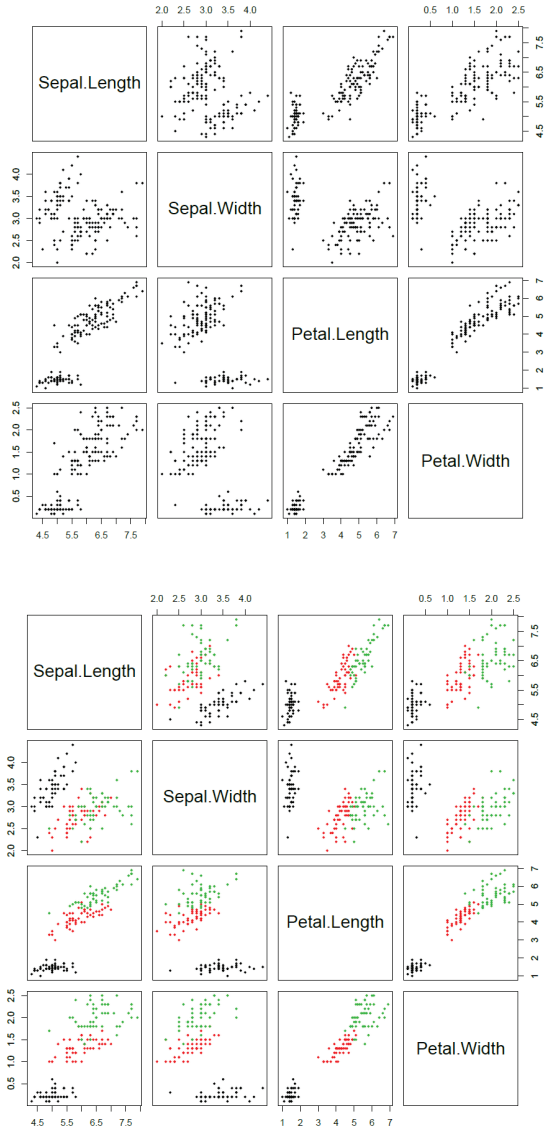

Figure 1: Pairsplots of the sepal length, sepal width, petal length, and petal width from Fisher's iris data, without and with consideration of iris species (measurements in centimeters). Black, red, and green observations are from species setosa, versicolor, and virginica respectively.

3 Statistical Clustering

Hierarchical clustering is a method that is motivated by finding the optimal step at each stage in the progressive subdivision or synthesis of the data, where

each step operates on a proximity matrix of some kind. Hierarchical clustering techniques may be subdivided into agglomerative methods, which proceed by a series of successive fusions of the n observations into clusters, and divisive methods, which separate the n observations successively into finer clusterings (Everitt et al. [2011]). Agglomerative hierarchical clustering is most commonly used. In this method, each observation begins as its own singleton cluster and the clusters are sequentially merged into larger clusters, until all elements are in one final cluster. Deciding on which observations/clusters to combine requires a metric space and linkage method. Both of these choices ultimately impact the final clustering solution.

The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. The choice of metric should reflect how one wants to define similarity between observations or clusters, where Euclidean and Manhattan distances are most commonly used. Define vectors $\mathbf{p} = (p_1, p_2, ..., p_n)$ and $\mathbf{q} = (q_1, q_2, ..., q_n)$. Then the Euclidean distance between $\mathbf{p}$ and $\mathbf{q}$ is defined as $d_E(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$, and the Manhattan distance is defined as $d_T(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|$ (Daepp and Gorkin [2011]). Euclidean distance is our typical notion of distance, described often by "as the crow flies" distance, and Manhattan distance is often referred to as the "city block" distance, because a grid-based path must be followed.

The linkage criterion (used with the chosen metric) determines the distance between (sets of) observations as a function of the pairwise distance between observations. The differences among linkage methods come from the definition of the links, namely the distance between observations from two distinct clusters. In single-linkage, the distance between two clusters is defined as the shortest distance between observations in each cluster. Single linkage is useful in outlier identification, but it often creates "chaining," where clusters are combined in a sequential, chain-like fashion. In complete-linkage, the distance between two clusters is defined as the longest distance between observations in each cluster. Clusters found from complete linkage are often compact and of the same size. In average-linkage, the distance between two clusters is defined as the average distance between observations in one cluster to every object in the other cluster. Average linkage is a compromise between single and complete linkages. It joins clusters with small variances and is robust against outliers. For all three linkages, two observations/clusters are merged based on the shortest distance (most similar observations/clusters) calculated at each step, where distance is defined by the choice of metric.

The biggest drawback of hierarchical clustering is that the method does not produce a unique partition of data, but rather a hierarchy from which the user needs to choose an appropriate number of clusters. A great advantage of the method is that regardless of the dimensionality of the data, the results of a hierarchical clustering solution are easy to visualize with a tree-like diagram called a dendrogram (see Figure 2). In a dendrogram, each observation is represented by a node at the bottom of the tree. In agglomerative fashion, the most similar observations are linked and this process continues up the tree until all of the observations are in one large cluster. The vertical distance between

branches is proportional to the similarity of the merged clusters, where a greater vertical distance indicates less similarity. Visually, a reasonable "cut point" for the tree, producing a final clustering solution, can be chosen by looking for a large vertical distance between branches of the dendrogram.

Hierarchical clustering with the two defined distances and three linkage methods are applied to the iris data using the `hclust` function in the statistical software `R`. Figure 2 shows an example dendrogram for hierarchical clustering with complete linkage and Euclidean distance. As with the pairsplot, the node color indicates the iris species. The dashed line provides a reasonable place to cut the dendrogram to obtain a final 3-group clustering solution. It can be seen that setosa species (black) is correctly clustered, but there is a mixture of the virginica and versicolor species across the other two clusters.

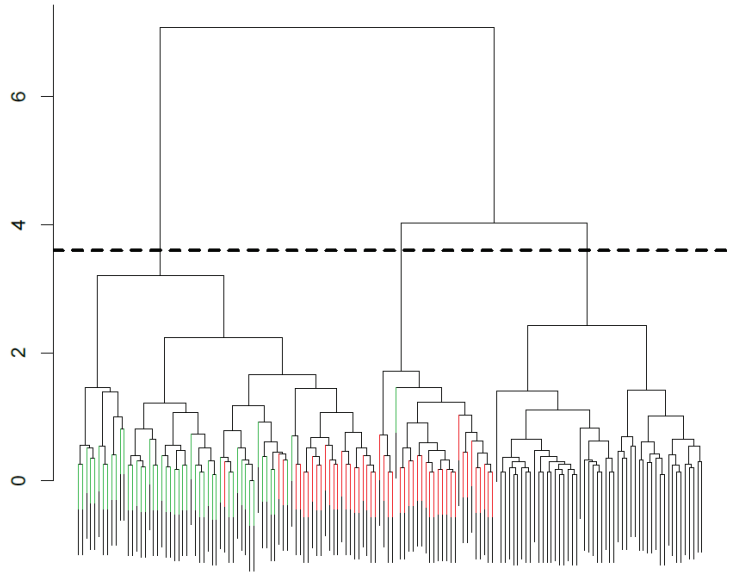

Figure 2: Dendrogram for hierarchical clustering with complete linkage and Euclidean distance of iris.

K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters, in which each observation is assigned to the cluster of the nearest centroid. K-means is a rather straightforward, well known algorithm for clustering observations. The main disadvantage of this algorithm is that the user must specify the number of clusters (k) they wish to identify prior to fitting the model. In practice, a range of values for k (usually 1 to some maximum number) are used and a final model is chosen based on some criterion (see Steinley [2006]).

For this method, each observation is thought of as being represented by some feature vector in a d-dimensional space, d being the number of all fea-

tures used to describe the objects to cluster. In that vector space the algorithm randomly chooses k observations as starting values for the "k means", namely $\mu_1^0, \mu_2^0, \ldots, \mu_k^0$, which serve as the initial centroids of the clusters. Next, typically using Euclidean distance, observations are assigned to the cluster with the nearest centroid. A new centroid is computed for each cluster by averaging the feature vectors of all assigned observations. The "k means" then update to $\mu_1^1, \mu_2^1, \ldots, \mu_k^1$. The process of assigning observations and recomputing centroids is repeated until there is a negligible change in cluster centroids. Specifically, if we let $\epsilon > 0$, then the convergence of the algorithm is defined as $\sqrt{\sum_i \left(\mu_i^{j+1} - \mu_i^j\right)^2} < \epsilon$, where $j = 0, 1, 2, \ldots$ (Bottou and Bengio [1995]). A k-means clustering solution is obtained by the final cluster assignment at convergence.

K-means clustering with 2, 3 and 4 clusters is applied to the iris data using the `kmeans` function in `R`. To visualize a k-means clustering solutions, a 2-dimensional projection of the data is often required. Using the `plot` function on a `kmeans` object will produce a visualization of the solution using the first two principal components (see Pearson [1901]). As an example, Figure 3 provides a 3 cluster solution projected from 4 dimensions down to 2. We see that cluster 2 (setosa species) is well separated from the other two clusters, while clusters 1 and 3 have overlap.

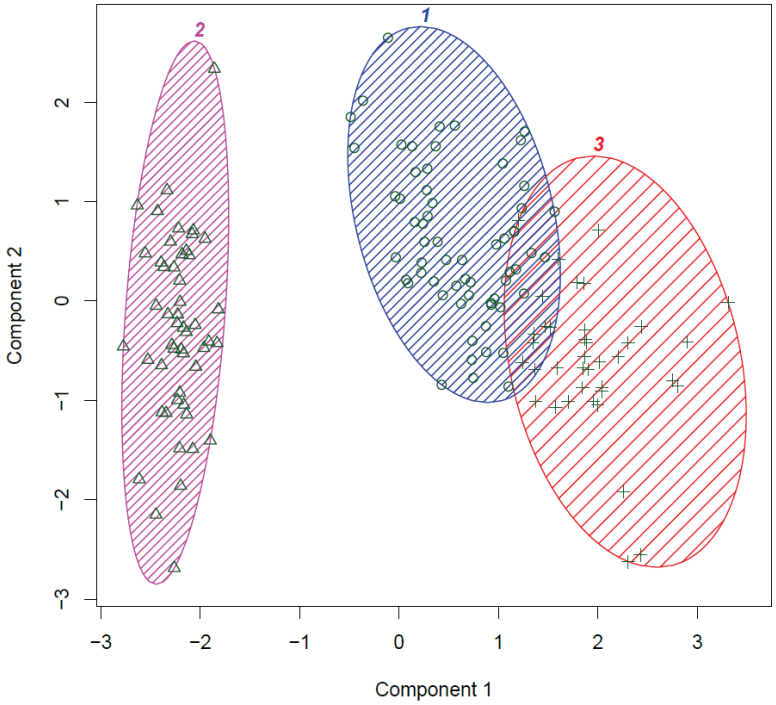

Figure 3: Two-dimensional principal component projection of the k-means clustering solution with k=3.

Model-based clustering aims to optimize the fit of a mathematical model to the observed data. Such methods are often based on the assumption that the data are generated by a finite mixture of underlying probability distributions (McLachlan and Peel [2000]). This mixture model of subpopulations is summarized by the following equation

$$f(x) = \sum_{k=1}^{K} \tau_k f_k(x),$$

where each $0 < \tau_k \leq 1$ and $\sum_{k=1}^{K} \tau_k = 1$. Here, f_k is the k^{th} component density in the mixture, often assumed to be Gaussian with corresponding mean and covariance matrix (μ_k, Σ_k), and τ_k is the proportion of the population in component k, where $k = 1, \ldots, K$. It is often assumed that each component or subpopulation in the mixture model maps to a cluster. The most widely used approach to fit this mixture model is the Expectation Maximization (EM) algorithm (Dempster et al. [1977]). In statistics, the EM algorithm is an iterative method for finding the maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent parameters (here, τ_k, μ_k, and Σ_k, for $k = 1, \ldots K$). The EM algorithm iterates between an expectation (E) step and a maximization (M) step. To begin, the unknown parameters are initialized for each cluster. This is often done by random assignment of observations to components, or by using a k-means solution. In the E-step, the log-likelihood is calculated and every observation is given a probability of belonging to each cluster, calculated using Bayes' Theorem. In the M-step, parameter estimates are updated based on the maximum of the log-likelihood estimates. The procedure iterates between steps until the mixture model converges (defined similarly to the convergence of the k-means algorithm).

Unlike k-means or hierarchical clustering, model-based clustering is parametric, meaning that a probability distribution must be chosen for model. The choice of distribution is essential and needs to reflect the structure of the data. For example, if the data has outliers, then a Gaussian mixture model will provide a poor fit and a mixture of t-distributions would be more appropriate. It is also of importance to note that unlike the other clustering methods, model-based clustering produces a soft assignment or partition of the data. That is, rather than being assigned to one cluster with complete certainty (hard assignment), an observation is given a vector of probabilities (soft assignment) whose entries estimate the observation's probabilities of belonging to each of the k clusters, respectively. These probability vectors can be made into hard assignments (e.g., for the purpose of applying agreement indices) by assigning an observation to the cluster with the highest probability of membership (known as maximum a posteriori estimation).

Estimating the covariance matrix Σ_k for high dimensional data can be extremely difficult. In some cases it is impossible. To reduce the number of parameters that need to be estimated, the `mclust` package in `R` fits different parameterizations of Σ_k. The 14 possible parameterizations are summarized in Table 1 (Fraley et al. [2012]), where the name is indexed by three let-

ters: the first position indicates the volume of the components, the second the shape, and the third the orientation. The possible options for these positions include 'E' for equal, 'V' for variable, or 'I' for identity matrix. While these parameterizations are strictly for estimating the covariance matrix in Gaussian mixture models, similar reduced parameterizations exist for mixtures of other distributions.

Name	Interpretation
"EII"	spherical, equal volume
"VII"	spherical, unequal volume
"EEI"	diagonal, equal volume and shape
"VEI"	diagonal, varying volume, equal shape
"EVI"	diagonal, equal volume, varying shape
"VVI"	diagonal, varying volume and shape
"EEE"	ellipsoidal, equal volume, shape, and orientation
"EVE"	ellipsoidal, equal volume and orientation
"VEE"	ellipsoidal, equal shape and orientation
"VVE"	ellipsoidal, equal orientation
"EEV"	ellipsoidal, equal volume and equal shape
"VEV"	ellipsoidal, equal shape
"EVV"	ellipsoidal, equal volume
"VVV"	ellipsoidal, varying volume, shape, and orientation

Table 1: Covariance parameterizations for fitting Gaussian mixture models in `mclust`.

One benefit to using model-based clustering is that the method is built on a mathematical model that can be optimized. Therefore, the number of clusters k can be chosen by some information criterion. The Bayesian Information Criterion (BIC, Schwarz [1978]) is most commonly used as it balances the maximum number of parameters needed for the model (a result of the parameterization) and overfitting the data, i.e., choosing too many clusters (Fraley and Raftery [2002]). Note that the likelihood of model $\mathcal{M}$ can always be increased as the number of clusters k tends to the number of observations n. Using a criterion like the BIC to choose the number of clusters helps balance the complexity of the model with the overall fit. For each parameterization of the covariance matrix, `mclust` computes the BIC for the mixture model as follows

$$BIC(M) = 2\log(\text{maximized likelihood of model } \mathcal{M}) - \nu\log(n),$$

where ν is the number of independent parameters in model $\mathcal{M}$. Then the number of components/clusters that maximizes the BIC score is considered the best clustering solution.

The BIC for model-based clustering solutions of the iris data are visualized in Figure 4. The horizontal axis provides the number of components fit (k)

and the vertical axis provides the BIC value. Each line represents a different parameterization of the covariance matrix. The 2 cluster "VEV" (ellipsoidal and equal shape) model has the maximum BIC of -561.73, among all models fit. The 3-cluster solution for this same parameterization has a nearly identical BIC value of -562.55.

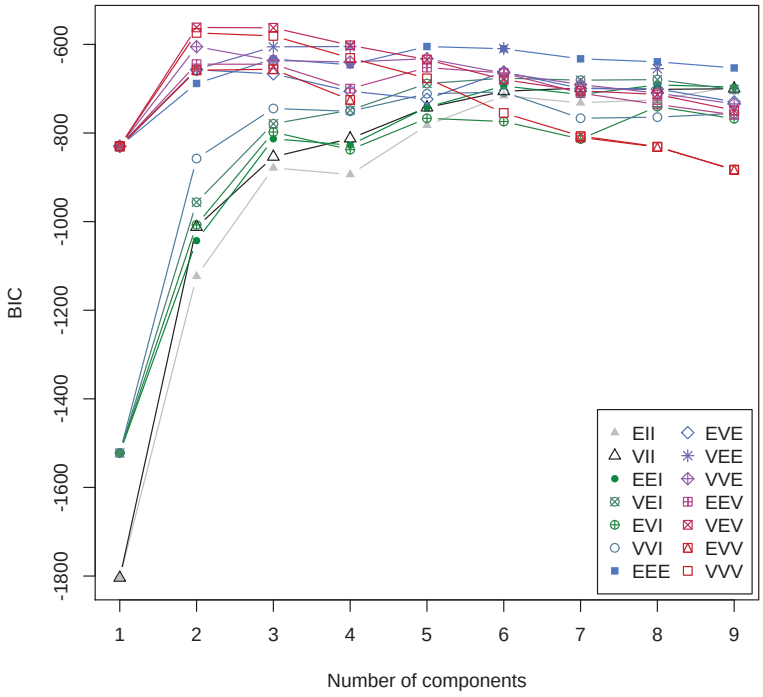

Figure 4: BIC values for of model-based clustering solutions of 'iris'.

4 Agreement Indices

As discussed in the introduction, a clustering solution summarizes the data by assigning class labels or cluster assignments to each observation. This produces a hard partition of the data, in which each observation is assigned to a single cluster with complete certainty. To compare clustering solutions to each other, or to a set of true class assignments (if known), we need a measure for evaluating the similarity of two partitions of the same data set. Agreement or performance indices are measures of correspondence between two hard partitions of the same data. There are numerous different agreement indices that all measure similarity in different ways; typically, however, the higher the value of the index, the greater the similarity between the partitions (Hubert and Arabie [1985]).

We will focus on three of the most commonly used agreement indices: the adjusted Rand, Fowlkes-Mallows, and Jaccard indices to compare the estimated clustering solutions to the "true" clustering solution of the iris into their species.

To understand how these three performance indices are defined and calculated, consider two partitions P and Q of a dataset which contains n observations. Suppose partition P has r clusters $p_1, \ldots, p_r$ and partition Q has c clusters $q_1, \ldots, q_c$. Now let n_{ij} be the number of observations that are in both cluster p_i and cluster q_j, for $i \in \{1, \ldots, r\}$, $j \in \{1, \ldots c\}$. Then $n_{i.} = \sum_{j=1}^{c} n_{ij}$ and $n_{.j} = \sum_{i=1}^{r} n_{ij}$ give the number of observations in cluster p_i and cluster q_j respectively. This notation is summarized in the contingency table provided in Table 2.

P \ Q	q_1	q_2	...	q_c	Sums
p_1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
p_2	n_{21}	n_{22}	...	n_{1c}	$n_{2.}$
...	...	...	...	...	...
p_r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
Sums	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	n

Table 2: Notation for the number of common observations in clusters between two partitions P and Q.

Further, we can define a as the number of pairs of objects in the same cluster in P and the same cluster in Q, b as the number of pairs of objects in the same cluster in P but different clusters in Q, c as the number of pairs of objects in the same cluster in Q but different clusters in P, and d as the number of pairs of objects that are in different clusters in P and in Q. This notation is summarized in Table 3. Note that a and d represent agreements and b and c disagreements.

$P \backslash Q$	Same Cluster	Different Clusters	Sums
Same Cluster	a	b	$a+b$
Different Clusters	c	d	$c+d$
Sums	$a+c$	$b+d$	$n=a+b+c+d$

Table 3: Describing cluster assignment of pairs of observations between two partitions P and Q.

Relating the notation in Table 2 to that of Table 3, we have the following:

$$a = \sum_i \sum_j \binom{n_{ij}}{2}, \quad b = \sum_i \binom{n_{i.}}{2} - a, \quad c = \sum_j \binom{n_{.j}}{2} - a, \quad d = \binom{n}{2} - a - b - c.$$

An advantage to counting pairs of observations in the same clusters, rather than individual observations, is that you can still find some agreement when a solution splits a cluster into finer clusters, or when a solution combines clusters into a larger cluster.

The Rand index (RI) is calculated by $\frac{a+d}{\binom{n}{2}}$ and is bounded between 0 and 1. In practice, the RI does not span its range and will only approach 1 as the number of clusters increases (Steinley [2004]). Furthermore, the RI does not have a closed form for its expectation. Due to these limitations, several extensions of the RI have been proposed.

The adjusted Rand index (ARI), proposed by Hubert and Arabie corrects the Rand index for chance by subtracting off the expected value of the index and dividing by the maximum value minus the expectation (Hubert and Arabie [1985]). The the maximum value of the ARI is still 1, indicating perfect similarity between the 2 partitions, and the expected value under random partitioning is 0 (Steinley [2004]). In instances of extreme dissimilarity between two partitions, the ARI will be negative. The ARI is defined as

$$\begin{aligned} ARI &= \frac{Index - Expected[Index]}{Max[Index] - Expected[Index]} \\ &= \frac{\sum_i \sum_j \binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} / \binom{n}{2}}{\frac{1}{2}[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}] - \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} / \binom{n}{2}} \\ &= \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \\ &= \frac{a+d-C}{a+b+c+d-C}, \\ &\quad \text{where } C = \frac{(a+b)(a+c)+(c+d)(b+d)}{\binom{n}{2}}. \end{aligned}$$

The Jaccard index is defined for any two sets (or partitions) as the size of the intersection divided by the size of the union, and is defined as

$$Jaccard = \frac{|P \cap Q|}{|P \cup Q|} = \frac{a}{a+b+c}.$$

It is easy to see that the Jaccard Index is bounded between 0 (no similarity) and 1 (perfect similarity).

The Fowlkes-Mallows index is another index (Fowles and Mallows [1983]) bounded between 0 and 1 that measures the agreement between two partitions as follows

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}.$$

To get a better understanding of the above indices, consider the example shown in Table 4.

P \ Q	q_1	q_2	q_3	Sums
p_1	1	1	0	2
p_2	1	2	1	4
p_3	0	0	4	4
Sums	2	3	5	10

Table 4: Illustrative example for calculating agreement indices.

Counting the pairs of common observations in the same clusters in two clustering solutions P and Q according to Table 3, we have $a = \binom{2}{2} + \binom{4}{2} = 7$, $b = \binom{2}{2} + \binom{4}{2} + \binom{4}{2} - 7 = 6$, $c = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} - 7 = 7$, and $d = \binom{10}{2} - 7 - 6 - 7 = 25$. Then we can calculate the performance indices to evaluate the similarity of partitions P and Q:

$$ARI = \frac{7 - 14 \cdot 13/45}{\frac{1}{2}(14 + 13)/45} = 0.313, \qquad Jaccard = \frac{7}{7 + 6 + 7} = 0.350,$$

$$\text{and } FM = \sqrt{\frac{7}{7+6} \cdot \frac{7}{7+7}} = 0.519.$$

Typically, the values of these indices are not compared, because they are measuring similarity of partitions in different ways. It is worth noting, however, that the Fowlkes-Mallows index will always be greater than or equal to the Jaccard index:

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} \geq \sqrt{\frac{a}{a+b+c} \cdot \frac{a}{a+c+b}} = \frac{a}{a+b+c} = Jaccard.$$

Through our application of these indices to data, we believe that the Fowlkes-Mallows index will also be greater than or equal to the ARI, while there will not be a particular relationship between the ARI and the Jaccard index.

5 Evaluating Clustering Solutions

Now we can measure the agreement between each clustering solution and the true classification of species using each agreement index. Tables 5 and 6 summarize the values of the agreement indices for hierarchical clustering solutions with Euclidean and Manhattan distance respectively. In both of these tables, as expected, the Fowlkes-Mallows index is larger than the ARI and Jaccard indices for all linkage methods and number of clusters, while the ARI and Jaccard indices are not significantly different. For single linkage, the number of clusters in the solution does not have an impact on any of the agreement indices. This is likely because single linkage is often characterized by "chaining" and so the addition of a new cluster, often means a single observation is added to a cluster. For complete linkage, the 3-cluster solution has greater similarity with the

species classification (regardless of agreement index used) than the 2-cluster or 4-cluster solutions. For average linkage, the 3-cluster solution has a slightly larger value for all three performance indices than the 4-cluster solution. In comparing the indices across distance metrics (Tables 5 and 6), there are small differences for single and average linkages. Complete linkage, however, offers the largest differences, where the 2-cluster solution using Euclidean distance is better at recovering the true classifications than Manhattan distance, but the opposite is true for the 3-cluster and 4-cluster solutions.

To summarize, solutions using Manhattan distance have greater similarity with true species assignment than those using Euclidean distance for single and complete linkages. Average linkage overall tends to produce more accurate clustering solutions, which are not significantly impacted by the choice of distance metric. Overall the 3-cluster hierarchical clustering solution produced using Euclidean distance and average linkage is most similar to the true species classification of the iris. It is also worth noting that the number of clusters producing the optimal solution is consistent across all three agreement indices, for any combination of linkage and distance metric.

Method	Number of Clusters	ARI	Fowlkes-Mallows	Jaccard
Single	2	0.568	0.771	0.595
	3	0.564	0.764	0.589
	4	0.562	0.760	0.586
Complete	2	0.422	0.665	0.482
	3	0.642	0.769	0.622
	4	0.589	0.720	0.562
Average	2	0.568	0.771	0.595
	3	0.759	0.841	0.725
	4	0.729	0.818	0.692

Table 5: Agreement indices comparing hierarchical clustering solutions with Euclidean distance to the iris species.

Table 7 shows the values of the agreement indices for k-means clustering solutions. For this clustering method, the choice of agreement index impacts which k produces the optimal clustering solution, and the 3-cluster solution is never optimal. The ARI indicate the most similarity in the 2-cluster solution, whereas the Fowlkes-Mallows index is greatest for the 4-cluster solution. The Jaccard index finds equivalent similarity among the 2-cluster and 4-cluster solutions. Recall that a 2-cluster solution is reasonable given the overlap between the versicolor and virginica species. The 4-cluster k-means solution separates the distinct setosa species into two smaller clusters.

Method	Number of Clusters	ARI	Fowlkes-Mallows	Jaccard
Single	2	0.568	0.771	0.595
	3	0.566	0.767	0.592
	4	0.564	0.763	0.589
Complete	2	0.245	0.607	0.403
	3	0.732	0.824	0.700
	4	0.656	0.761	0.608
Average	2	0.568	0.771	0.595
	3	0.745	0.831	0.710
	4	0.715	0.808	0.678

Table 6: Agreement indices comparing hierarchical clustering solutions with Manhattan distance to the iris species.

k	ARI	Fowlkes-Mallows	Jaccard
2	0.616	0.733	0.573
3	0.433	0.664	0.485
4	0.540	0.750	0.572

Table 7: Agreement indices comparing k-means clustering solutions to the iris species.

Finally, Table 8 shows the agreement indices for model-based clustering solutions of the "VEV" parameterization of the iris data. Though the BIC selects the 2-cluster "VEV" model as the optimal clustering solution, all three performance indices show the most similarity with the true species classifications for "VEV" with three clusters. The values of the agreement indices are very close to 1, indicating that these are the most similar partitions.

Number of Clusters	ARI	Fowlkes-Mallows	Jaccard
2	0.568	0.771	0.595
3	0.904	0.936	0.879
4	0.805	0.867	0.763

Table 8: Agreement indices comparing the "VEV" model-based clustering solutions to the iris species.

The most accurate clustering solutions for recovering species for Fisher's famous iris data are for hierarchical clustering, the 3-cluster solution with average linkage and Euclidean distance; for k-means clustering, the solution with

2 or 4 clusters; and for model-based clustering, the 3-cluster, "VEV" parameterization. Among these three, model-based clustering produced the solution the highest overall agreement with the true iris species classification.

6 Summary

This paper introduced three of the most commonly used clustering methods: hierarchical clustering, k-means clustering, and model-based clustering; as well as three performance indices: the adjusted Rand index, the Jaccard index and the Fowlkes-Mallows index. The three clustering methods were used to cluster the iris dataset and their performances were measured with all three indices. After comparing the performance indices of the various clustering solutions, it is concluded that model-based clustering with 3 clusters (model 'VEV') produces a clustering solution that best recovers the true classification of iris into species.

A comparison of this nature is only possible because we can compare our clustering solution with "true" known labels. Clustering is an exploratory technique and when performing these methods on data without a true classification, it is not always clear which method will be best. A simulation that closely resembles the real data can help a researcher choose the most appropriate method. Additionally, the type of data can help dictate the method - for example, if there is a nested structure, then one would naturally choose hierarchical clustering.

Agreement indices can also be used to compare the similarity of clustering solutions produced for different numbers of clusters, using the same method or across methods as an additional way to understand the relationships of the methods and indices. While we have presented work on the most common clustering methods and agreement indices, there are many others left to be explored.

Bibliography

[1] Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988) *The new S language*. Wadsworth & Brooks/Cole.

[2] Bottou, L., and Bengio, Y. (1995) Convergence properties of the k-means algorithms. *Advances in Neural Information Processing Systems 7*. The MIT Press.

[3] Daepp, U., and Gorkin, P. (2001) *Reading, writing, and proving: a closer look at mathematics*. Springer.

[4] Dalgaard, P. (2008) *Introductory Statistics with R (Statistics and Computing)*. Springer.

[5] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. Series B (Methodological), 39(1):1–38. Wiley-Blackwell.

[6] Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011) *Cluster analysis*. Wiley.

[7] Fowlkes, E. B., and Mallows, C. L. (1983) A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*. 78(383):553. American Statistical Association.

[8] Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012) Mclust version 4 for r: normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report 597*, Department of Statistics, University of Washington.

[9] Fraley, C., and Ratery, A. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, (97):611–631. American Statistical Association.

[10] Hubert, L., and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, (2):193–218. Springer.

[11] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979) *Multivariate analysis Probability and mathematical statistics*. Academic Press.

[12] McLachlan, G. J. and Peel, D. (2000) *Finite mixture models*. Wiley.

[13] Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. 2(11):59–572. Taylor & Francis.

[14] Schwarz, G. E. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464. Institution of Mathematical Statistics.

[15] Steinley, D. (2004) Properties of the Hubert-Arabie adjusted rand index. *Psychological Methods*, (9):386–396. American Psychological Association.

[16] Steinley, D. (2006) K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34. Wiley-Blackwell.

[17] Yeung, K. Y., and Ruzzo, W. L. (2001) Details of the adjusted rand index and clustering algorithms, supplement to the paper "an empirical study on principal component analysis for clustering gene expression data". *Bioinformatics*, 17(9):763–774. Oxford University Press.