Bayesian Hierarchical Clustering

Nicholas Lowing & Ryan Bomalaski
Group 3
CSE 5290
Dr. Debasis Mitra

April 28, 2017

Introduction

In this project Bayesian Hierarchical Clustering (BHC) was reviewed. A paper by Heller et al.[1] was used as the primary source of information on the BHC algorithm. Multiple software packages implementing different types of Bayesian Clustering were obtained. The BHC software package[2], written in R/C++ and modeled on Heller et al.[1], was studied in the most detail and applied to data sets. Overall, BHC was successful in clustering data according to predictions for Fisher's Iris data and a Zoo Animal data with accuracies in excess of 95%.

Overview of Clustering

Clustering is an important technique for grouping data elements and identifying underlying patterns. This section provides an overview of clustering and its significance, as well as an in-depth view of the BHC algorithm courtesy of Heller et al.[1]

Clustering is an analysis technique that groups data elements into clusters. It is mainly used as a technique for preprocessing data prior to the application of advanced machine learning algorithms. These clusters provide a judgment of the similarity of individual data elements amongst the entire data set. In this sense, the goal of clustering is to produce clusters of data elements of high intracluster similarity; that is, data elements within a single cluster are most closely related. The determination of *closely related* is based upon the clustering algorithm.

There are many types of clustering algorithms that can be classified by how the concept of the *cluster* is defined. BHC is an example of a connective model, where clusters have a distance definition. K-means is another popular type of clustering algorithm, which is a centroid-based model that uses a mean vector to represent a cluster. There are many others,

all of which differ on how the clusters are formed. In this project, *hierarchical* clustering was the focus. This means that all clusters contain data elements that have high intracluster similarity, clusters are based on a distance measure, and all clusters are also placed within a nested structure. All data elements belonging to two different child clusters that are merged also belong to the parent cluster. An example dendrogram (hierarchical structure) is shown in **Fig. (1)**.

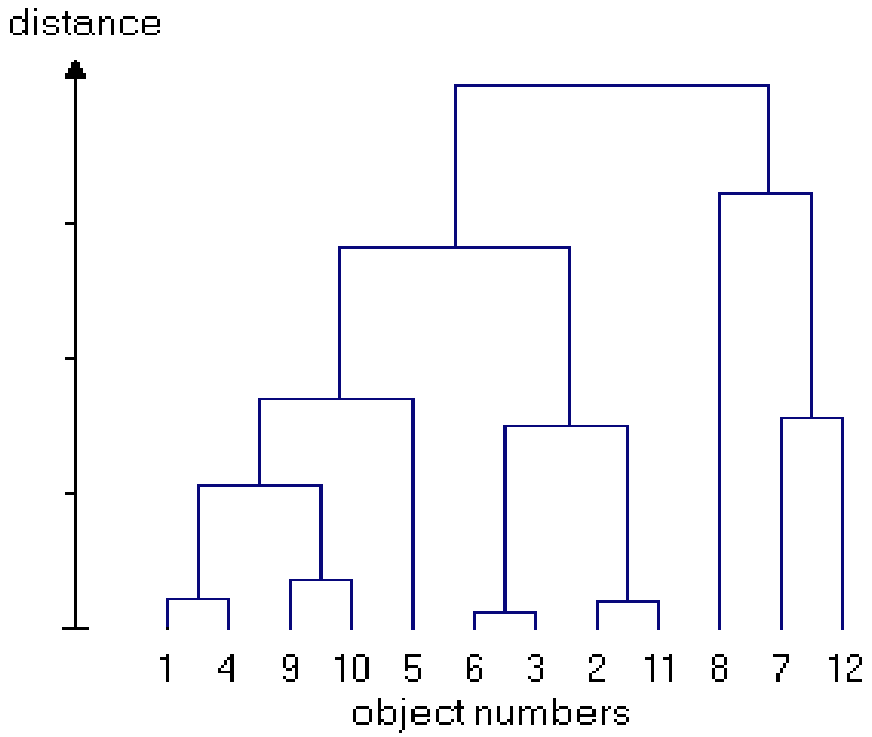

Fig. 1: *An example dendrogram which is the hierarchical clustering algorithm's visual output for a data set.*

Further distinction can be made regarding connective clustering algorithms by analyzing the distance measure used. Euclidean distance, Manhattan distance, etc... are all types of physical distance measures. For BHC, a probabilistic distance measure is used. This means that rather than a fixed distance determining clustering, the *distance* is actually the probability that data elements should be present in a particular cluster. This becomes important in determining on clusters iteratively merge to form new clusters within the hierarchical structure. *Bayesian* refers to the posterior probability calculation that is used to define the algorithm's distance measure.

Bayesian Hierarchical Clustering (BHC)

This section will discuss the BHC algorithm as presented in Heller et al.[1].

Overview of Algorithm

BHC is a one-pass, agglomerative, hierarchical clustering algorithm that uses a Bayesian probabilistic distance measure. This means the algorithm is greedy and clusters in a bottom-up manner. Bottom-up refers to how clusters are formed within the hierarchy. All data elements are initialized within individual clusters and then clusters are iteratively merged until no more mergers occur. Mergers are based upon a posterior probability calculation that makes use of Bayes Theorem. The final result is a dendrogram that shows a hierarchical structure of the input data set.

The BHC algorithm presented in Heller et al.[1] is shown in **Fig. (2)**. The algorithm can be broken down into the following steps:

(1) Initialize all data elements into individual clusters (represented by nodes of a tree) with probability arrays. Select the model and prior that will be used.

(2) Compute the probability arrays for each cluster. First element of the array is the probability that the cluster remains independent. All other elements of the array are the probabilities of merger with all other clusters.

(3) Merge clusters based on the highest probability within each cluster's probability array.

(4) Create the merger cluster and assign it as the parent of the two child clusters. This includes computing the probability array for the merger (parent) cluster.

(5) Terminate the algorithm when no more mergers occur.

The algorithm will produce a dendrogram of the resulting hierarchical structure, allowing for a visual analysis of the clustering outcome. The mathematical symbolism and formulation for the algorithm will be presented and discussed further in the next section.

input: data $\mathcal{D} = \{\mathbf{x}^{(1)} \ldots \mathbf{x}^{(n)}\}$, model $p(\mathbf{x}|\theta)$, prior $p(\theta|\beta)$
initialize: number of clusters $c = n$, and $\mathcal{D}_i = \{\mathbf{x}^{(i)}\}$ for $i = 1 \ldots n$
while $c > 1$ **do**
Find the pair $\mathcal{D}_i$ and $\mathcal{D}_j$ with the highest probability of the merged hypothesis:

$$r_k = \frac{\pi_k p(\mathcal{D}_k|\mathcal{H}_1^k)}{p(\mathcal{D}_k|T_k)}$$

Merge $\mathcal{D}_k \leftarrow \mathcal{D}_i \cup \mathcal{D}_j, \quad T_k \leftarrow (T_i, T_j)$
Delete D_i and D_j, $c \leftarrow c - 1$
end while
output: Bayesian mixture model where each tree node is a mixture component
The tree can be cut at points where $r_k < 0.5$

Fig. 2: *BHC algorithm as presented in Heller et al.*[1]

Algorithm Formulation

Let $D = \{x^{(1)}, ..., x^{(n)}\}$ represent the data set of n elements. $D_i \subset D$ is the set of data elements belonging to cluster i, which is located in the leaves of subtree T_i. Initialization of the algorithm creates n clusters, $\{T_i : i = 1, ..., n\}$, which each contain a single data element. The algorithm proceeds by iteratively merging clusters. This is done by combining subtrees T_i and T_j into a new tree T_k where the set of data elements in T_k is $D_k = D_i \cup D_j$. This subtree concept is shown in **Fig. (3)**.

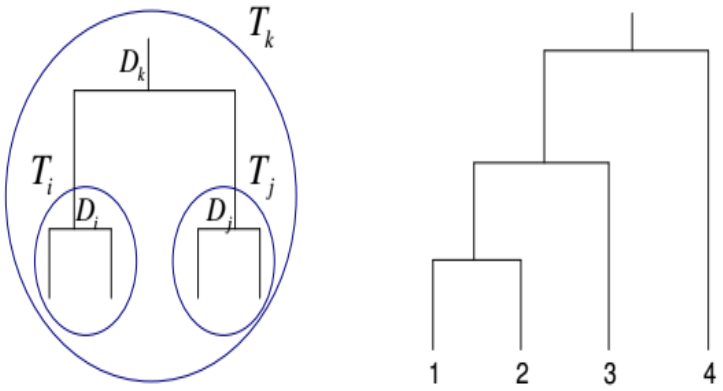

Fig. 3: *Subtree formulation used in the BHC algorithm presented in Heller et al.*[1] that produces a hierarchical clustering structure.

For each merger, there are two hypotheses that are compared. The first, H_k^1, is that all the data elements belonging to D_k were generated independently and identically from the same probabilistic model $p(x|\theta)$, where θ are unknown parameters. An example is the normal Gaussian distribution, which has mean and covariance parameters $\theta = (\mu, \Sigma)$. The model selected needs to be appropriate for the type of data being processed. In addition to a model, there is the need for a prior $p(\theta|\beta)$, where β are hyperparameters. This allows for the data under hypothesis H_1^k to be computed (assuming the prior is a *conjugate prior* of the probabilistic model) as

$$p(D_k|H_1^k) = \int p(D_k|\theta)p(\theta|\beta)d\theta \tag{1}$$

The second (alternative) hypothesis, H_2^k, is that the data subset corresponding to D_k has two or more clusters in it. The probability of the data under hypothesis H_2^k is just the product of the subtrees

$$p(D_k|H_2^k) = p(D_i|T_i)p(D_j|T_j) \tag{2}$$

Let the prior that all points in D_k belong to a single cluster be denoted by $\pi_k \equiv p(H_1^k)$. Both hypotheses can be combined in order to compute the marginal likelihood of data in a given tree T_k

$$p(D_k|T_k) = \pi_k p(D_k|H_1^k) + (1 - \pi_k)p(D_i|T_i)p(D_j|T_j) \tag{3}$$

Let the posterior probability of the merger hypothesis, $H1^k$, be denoted by r_k. Bayes Theorem is given by

$$p(B|A) = \frac{p(A)p(A|B)}{p(B)} \tag{4}$$

Applying **Eq. (4)** yields the posterior probability for the merged hypothesis

$$r_k = \frac{\pi_k p(D_k|H_1^k)}{p(D_k|T_k)} \tag{5}$$

A probability $r_k \geq 0.5$ suggests that a merger between clusters is more likely than not.

BHC Application and Results

The BHC software package implements the algorithm presented in Heller et al.[1] using a Bernoulli probabilistic model and a beta distribution conjugate prior. This section will give an overview of the data sets used for testing the BHC software package, as well as the dendrogram and accuracy results for each. BHC is available from the Bioconductor repository.[2]

Fisher's Iris Data Set

The first data set examined was Fisher's Iris Data Set - a set of iris flower data introduced by biologist Ronald Fisher in his paper *The use of multiple measurements in taxonomic problems*[3]. The data set contained values for 150 iris flower - 50 each for the species *Iris virginica*, *Iris versicolor* and *Iris setosa*. For each flower, measurements on sepal length, sepal width, pedal length and pedal width were recorded. A field was also included for each flower labeling its species. The Fisher Iris data set is available from the University of California, Irvine Machine Learning Database.[4]

Historically, this data set has not been used for cluster analysis, as it creates two obvious clusters and tends to not sub divide those clusters as desired. This is most readily seen when performing k-means clustering on the data. However, since BHC seeks to create hierarchical clusters reminiscent of phylogenetic trees, the data set held value as a test case.

Two versions of the iris data set were used. In the first instance, species label data was left in. This was done to provide a guide for the algorithm and allow confirmation of results. These results are show in **Fig. (4)**.

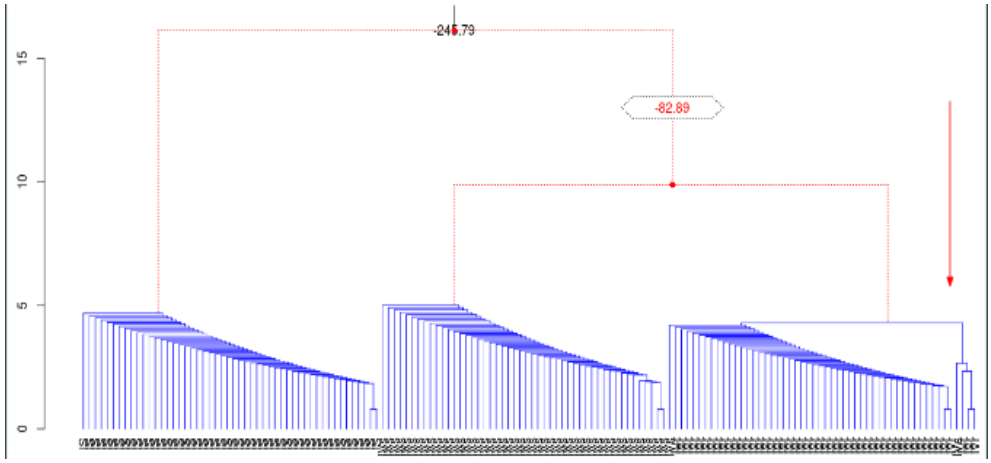

Fig. 4: *BHC dendrogram results using Fisher's Iris Data Set with species labels included. The red arrow denotes the 1 instance of incorrect grouping.*

The resulting data showed that, given the label data, BHC was able to handle the iris data set in a way other clustering algorithms were previously not capable of. Of the 150

cases, only 1 was incorrectly grouped. This grouping occurred between the closely related *Iris versicolor* and *Iris virginica*. However, it could be argued that providing the species label data gives BHC a bias towards the proper categorization.

The second version of the iris data set used was with the species labels removed. This was to provide confirmation that BHC was able to correctly handle the difficult data set, and provide results as anticipated. These results are shown in **Fig.(5)**.

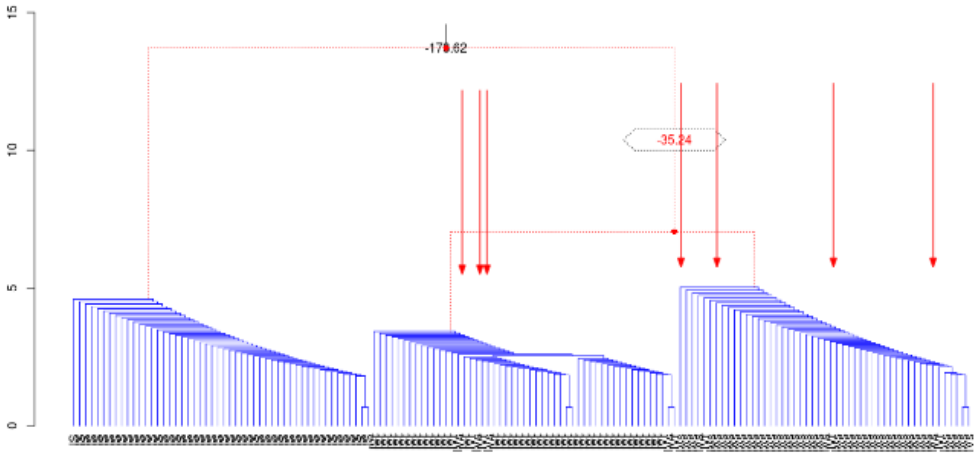

Fig. 5: *BHC dendrogram results using Fisher's Iris Data Set with out species labels included. The red arrows denote the 7 instances of incorrect grouping.*

Data from the second set showed that, even without the label data, novel results are still produced. The data showed proper grouping, and sub grouping, with *Iris setosa* set on a higher branch from *Iris virginica* and *Iris versicolor*. Though 7 of the flower samples were incorrectly grouped, BHC was still able to maintain above 95% accuracy.

Zoo Animal Data Set

The second data set used was Richard Forsyth's zoo data base. This data set contained 101 unique animals as entries, with 15 boolean descriptors and 2 numeric descriptors assigned to each entry. Each animal was given a true or false value for the following descriptors: hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic and catsize. Numerical values were assigned to the descriptors for legs and type, with legs being the count of legs and type assigning a label from 1 to 7 to select from: mammal, bird, reptile, fish, amphibian, insect, and other.

This dataset was chosen for two reasons. First, it provided easily verifiable data that BHC could process quickly. The data base was of information that a lay-reader could understand without requiring expertise in the field that the data represents. The second reason was that the data set provided label data, to allow for a labeled and unlabeled comparison as was done with Fisher's Iris Data Set. The results of the labeled and unlabeled analysis can be

seen in **Fig. (6)** and **Fig. (7)** respectively.

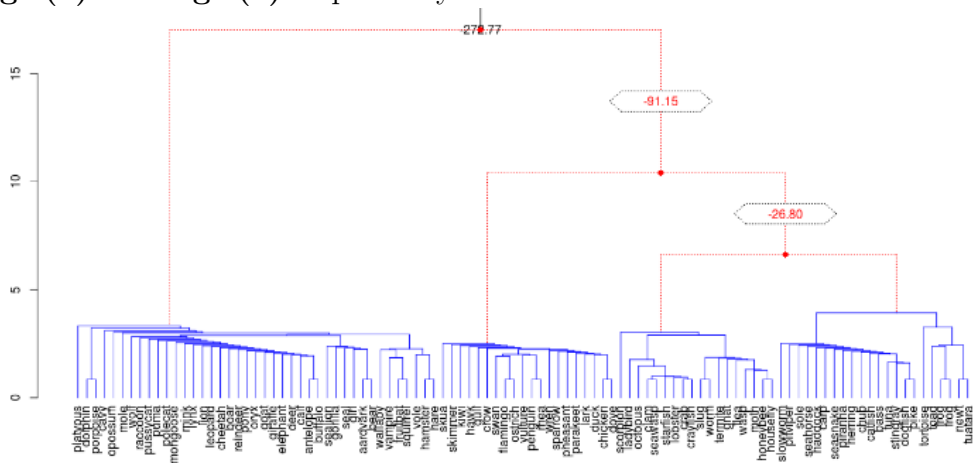

Fig. 6: *BHC dendrogram results using Forsyth's zoo data base with species labels included.*

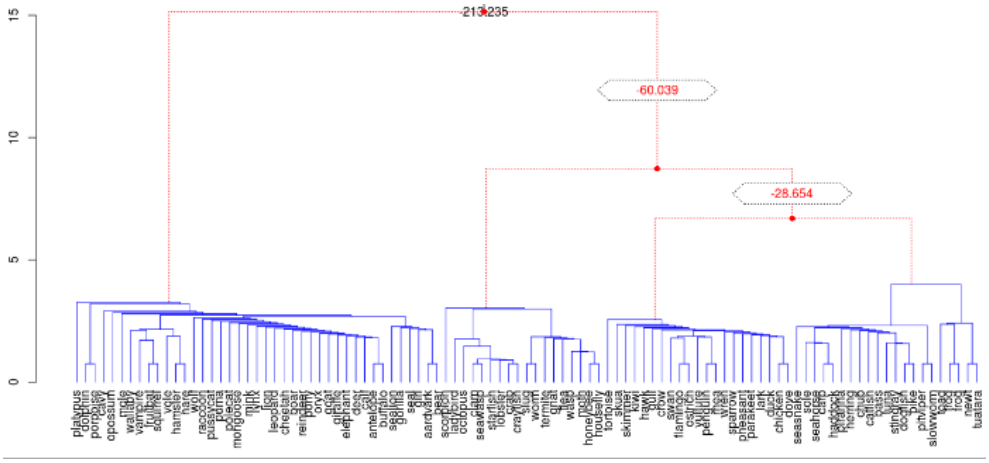

Fig. 7: *BHC dendrogram results using Forsyth's zoo data base without species labels included.*

In both data sets, clear animal class group distinctions were made as expected in both the labeled and unlabeled data. There were differences between the two dendrograms, though. Since labels were numerically represented, BHC inferred a closer relation between labels 3 and 4 than it would 3 and 6. Removing this label data shifted the large class branches without breaking up the sub classes. Thus, the large branching differences can be attributed to class labels being given these numeric weights. The Forsyth zoo data set is available from the University of California, Irvine Machine Learning Database.[4]

Summary

In summation, Bayesian Hierarchical Clustering has demonstrated that it is well worth its place on the tool belt of artificial intelligence and machine learning researchers. Heller et al.[1] have shown a novel algorithmic approach to BHC, which Savage et al.[2] have implemented in the BHC software package. Using the Fisher and Forsyth data sets, the BHC software package, and the BHC algorithm in general, was shown to work well with highly regarded machine learning data bases. The BHC software package was even able to handle a historically difficult data set.

Further research into Bayesian Hierarchical Clustering would be beneficial to further flesh out this cursory outline. The R repository bclust, while outside of the scope of this paper, was found to also work with a Bayesian Hierarchical Clustering algorithm.[5] Non-hierarchical Bayesian Clustering algorithms also exist, as do software with these algorithms implemented. Of particular note is the AutoClass family of software packages. AutoClass C is the open source implementation of this group of software. It was developed by Dr. Peter Cheeseman, John Stutz and Will Taylor for use by the National Aeronautics and Space Administration. It is currently maintained by Jams R. Van Zandt.[6]

Bibliography

[1] Heller, K.A., & Ghahramani, Z. (2005). Bayesian Hierarchical Clustering. *Proceedings of the 22nd international conference on machine learning*, pp.297-304. Retrieved from http://www2.stat.duke.edu/ kheller/bhcnew.pdf.

[2] Savage R., Cooke E., Darkins R. & Xu Y. (2011). *BHC: Bayesian Hierarchical Clustering*. R package version 1.28.0.Retrieved from https://www.bioconductor.org/packages/release/bioc/html/BHC.html.

[3] Fisher,R.A.(1936) *The use of multiple measurements in taxonomic problems* Annual Eugenics, 7, Part II, 179-188.

[4] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Partovi Nia, V. & Davison, A. (2015). *bclust: Bayesian Hierarchical Clustering Using Spike and Slab Models*. R package version 1.5.0. Retrieved from https://cran.r-project.org/web/packages/bclust/index.html.

[6] Cheeseman, P., Stutz, J., & Taylor, W. (2002) National Aeronautics and Space Administration *Autoclass - Automatic classification or clustering* (Version 3.3.6) Retrieved from https://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/autoclass-c/.