Supervised Hierarchical Clustering Using CART

T. P. Hancock[a], D. H. Coomans[a], Y. L. Everingham[a,b]

[a]Department of Mathematics and Statistics, James Cook University, Townsville, Queensland, Australia 4811
[b]CSIRO Sustainable Ecosystems, Davies Laboratory,Townsville, Queensland 4814, Australia

Abstract: The size and complexity of current data mining data sets have eclipsed the limits of traditional statistical techniques. Such large datasets frequently require some form of cluster analysis, usually in the form of a hierarchical cluster analysis. However the implementation of a traditional hierarchical scheme on large datasets requires an additional cluster validation analysis. Classification and Regression Trees (CART) are a non-parametric regression and classification technique that have become popular within the biotechnology and ecological fields. CARTs intuitive interpretation, and ability to handle large datasets make it easily accessible to the non-statistician by presenting the statistical relationships found in the form of a binary tree. This paper proposes a supervised clustering algorithm capable of finding real clusters within large datasets by using CART as a means of filtering the clusters found using any hierarchical technique. The supervision performed by CART acts as a filter of the results from a hierarchical cluster analysis by merging or removing poorly defined groups. It is common practice to validate a cluster analysis using descriminant analysis, however this assumes that the correct number of clusters is known. CART implements a selective classification of groups allowing for some groups not to be explicitly classified, a feature not supported by standard descriminant analysis. This selective classification acts in two fold, firstly by filtering or merging clusters that are not validated by the data, and secondly, as a relationship model for the clusters found and provides statistical measures of certainty over the analysis. An example of this method is presented using Sea Surface Temperatures (SST). This is an ideal choice as very little statistical cluster analysis has been implemented on this dataset, yet knowledge of such structure is in high demand. The analysis is performed for one month November for the years 1940 through to 2002, where some of the most useful variation is expected. The supervised clustering technique successful extracted seven meaningful clusters, which predicted with a cross-validated classification rate of 0.50.

Keywords: *CART, Supervised Clustering, Sea Surface Temperatures*

1. INTRODUCTION

Data mining sized datasets are now becoming the increasingly common in statistical research. The sizes of these datasets are large enough to confuse the standard statistical methods of analysis. Traditional hierarchical clustering was one of the first to collapse under the pressure of high dimensionality. However the question of *"How many clusters exist in a dataset?"* has become increasingly more important. Research fields such as biotechnology, medical and agricultural industries now produce datasets with in excess of 500 variables. Performing hierarchical clustering on such large numbers of variables results in a large numbers of potential clusters, but no real assurance of cluster validity.

Analysis of global sea surface temperatures (SST) is becoming an essential part of long-term climate forecasting models. As the trends within the SSTs are slow moving, any change will have a profound impact on the future climatic trends. Such information is particularly useful for the agricultural industry within Australia as knowledge of future climatic trends will lead to improved management decisions and improve the overall profitability and the stability of many industries.

Research by Drosdowsky (1993) provided a general study of the signals within the world's oceans. This research identified two independent signals within the Pacific and Indian Oceans respectively. Well known regions of importance to global climate trends are the Niño regions. These regions are located along the central and eastern equatorial Pacific. The important information on the structure and location of these regions can be used to further verify the validity of any cluster analysis.
Any cluster analysis of global sea surface temperatures (SST) variables will fall into the data-mining category. Although it is known that there are well-established clusters within the SSTs, there has been no explicit cluster analysis performed to try and extract these regions. The SST data comes in 180 by 360 degree images of the world, containing one measurement per degree, totalling to 64800 variables. For the purposes of this analysis it is

more suitable to examine the SST Anomalies (SSTA). These will be computed through subtraction between the raw SST data and the expected climatology, for this study the Reynolds SST climatology was used (Reynolds, 2001). As the SST data are interpolated surfaces the region between 40°E to 90°W longitude and 35°N to 55°S latitude is extracted as this region contains the most accurate information. Because the SST variables are highly correlated and this research is concerned only with large scale variability a 5-degree block average was computed over this region, with land and ice pixels removed resulting in 811 variables.

Classification and Regression Trees (CART), developed by Breiman et al. (1984), offer a unique statistical method for modeling a response. In this paper the classification algorithm is considered. This algorithm creates a binary tree from the set of predictor variables. The variables are selected such that the relative error (RE) statistic is minimized.

$$RE(d) = R_{LEFT}\left(d\right) + R_{RIGHT}\left(d\right) \qquad (1)$$

Where *RE* is the relative error, *d* is the decision rule that splits the primary node into left and right, R_{LEFT} is the risk of the left partition and R_{RIGHT} is the risk of the right partition.

At each node within the tree a decision rule, computed from the minimization of the error between that variable and the predictor, imposes a condition on each case. If the condition is satisfied, the tree is followed down the left sub-tree; otherwise the right sub-tree is evaluated. The risk for each new node is computed using the *'gini'* index (Brieman et. al., 1984).

$$GINI = \sum_i \sum_j L(i,j) p_i p_j \qquad (2)$$

Where *i* is the number of cases, *j* is the number of categories, *L(i,j)* is the loss of assigning case *i* to *j*, p_i is the probability of *i* and p_j is the probability of *j*.

Once an optimal tree is selected, by minimizing RE, the most populous category at each terminal node of the tree provides a classification of all the cases within that node. Classification trees have the ability to extract structure from large datasets with either categorical or numerical predictor variables, highlighting non-linear relationships between these predictor variables and the response. Additionally, bootstrapping and cross-validation methods are well developed within the tree framework, allowing for improved stability under prediction. The code used for this method is implemented in S-Plus using the TreesPlus (G. De'ath, 1999) module, which is a wrapper for the RPART C code (Thernau and Atkenson, 2000).

Hierarchical clustering (Everett, 1974) is a very common statistical technique for finding clusters within data. The algorithm implemented here is a complete linkage technique using euclidean distances. This technique was chosen as its observed performance appeared to generate more meaningful, larger clusters than other clustering methods tested. The code for this method is implemented by the *'hclust'* function in S-Plus 2000, Release 1 (1998).

2. SUPERVISED CLUSTERING ALGORITHM

The supervised clustering algorithm proposed is an iterative procedure, which first uses a hierarchical clustering technique to assign each variable into a group. Once the groups have been determined, a CART model is used to produce a correct classification rate (CCR), which assesses the performance of the clustering scheme. The classification strategy implemented by CART is one where the clusters that are easily classified, are classified first at the top of the tree. Therefore the order of classification down a tree is a ranking of how well each cluster is represented by the data. If the number of clusters found by the hierarchical scheme is initially set much higher than the expected number, it is possible to filter the clusters that are not well represented by the predictor variables. This process is done by simply selecting the tree size to be smaller than the number clusters found in the hierarchical technique.

At the end of each iteration, those variables that are misclassified by CART are removed, as they do not belong to the best subset of clusters. Once this is done the model can be made simpler by reducing the number of clusters to be found. This process is then iterated until the classification rate of the tree classifying the clusters has exceeded a preset tolerance. At this point the clusters that are left are those that are best represented by the data. This algorithm is more explicitly defined in Figure 1.

```
While CCR < tolerance.
1.  hierarchical_clusters = HCLUST(X,
    number_of_clusters)
2.  tree = RPART(Y = hierachical_clusters, X =
    $X^T$, Size = number_of_clusters)
3.  predicted_clusters = PREDICT(TREE =
    tree,X = $X^T$)
4.  CCR = SUM(predicted_clusters =
    hierarchical_clusters) / number_observations
5.  IF (predicted_cluster IS NOT EQUAL
    hierarchical_cluster)
        a.  remove that variable from the
            analysis
6.  number_of_clusters =
    round(number_of_cluster*CCR)

Where X is the data matrix of variables to be
classified, and CCR is the correct classification
rate.
```

Figure 1. Outline of the supervised clustering scheme

On the first iteration the hierarchical clustering should be specified such that it is an overfit of the data. The tree is then grown to classify these clusters based on the observations. To allow CART the chance to classify each cluster the size of the tree produced should be equal to or greater than the total number of clusters extracted. This is to ensure that each group has a chance to be classified. From here the tree is then used to predict the cluster membership of each variable. A classification rate is then computed, and the variables that have been classified into incorrect clusters are removed from the dataset. The number of clusters to be found in the next iteration is then reduced. This is a reduction in complexity of the clustering scheme that flows to a reduction CART tree size.

There are three main parameters that will affect the outcome of this technique. Firstly the classification tolerance must be chosen such that the performance of the tree classifying the clusters is at a maximum after all the iterations. For this reason, the tolerance level is set quite high; usually at a 90 to 95 percent CCR. Because the tolerance is specified externally, the CCR produced after the iterations are likely to be artificial. To determine how well the clusters are being classified, leave-one-out cross-validation, on the CCR of the tree (CV-CCR) will give a more realistic indication of overall model performance. The tolerance level with the highest CV-CCR will be selected for use in the final model.

The second parameter of importance is the size of the tree used to classify the clusters. Because the tree is constructed on the basis of which categories are classified best at each node, it follows that building a tree of size equal to the number of clusters does not mean each cluster will be explicitly classified. In the cases where some clusters are poorly defined, the tree classifies a subset of the best clusters. If this occurs a high rate of misclassification is expected. By filtering the misclassified variables, those clusters that well defined become clear. Because of the filtering process some uncertainty over the optimal size of the tree exists, because simpler tree could possible be generated to model the same clusters. The optimum number of clusters is found when the tree explicitly classifies each cluster within the hierarchical scheme. At this point the cross-validated performance of the tree is at a maximum. Therefore the final tree size, is dependant solely on the number of clusters within the model and can be determined on the basis its CV-CCR.

The poor classification performance of CART is the reason for removing those variables that are not correctly classified, rather than re-assigning them to their classified cluster or to a new misclassified cluster. When dealing with large numbers of highly correlated data, as in the SST dataset, CART can be confused as to which variable to make the split, as there can be many variables which find the same minimum RE. However the predictive performance of these variables within the final model will vary. Removing some of these variables is reducing the number of competitors for each split and thus increases the confidence in the final split. This in turn increases the overall performance of the tree. In this paper those variables that are removed during the supervision will be re-classified using the optimum tree found at the end of the algorithm.

The third important parameter is the specification of the initial number of clusters to be found. It is recommended that this number be a deliberate overfit to how many clusters are expected. This is performed because the order of extraction using hierarchical clustering is not necessarily the order of the most important clusters. In many cases, particularly when dealing with highly correlated data, the scheme will find outliers and smaller, less interesting clusters, first, rather than the most statistically valid clusters. Thus overfitting the number of clusters first will provide for some of the more interesting clusters to be identified through the filtering process.

A bi-product of this analysis is the predictive classification tree produced by CART. This tree allows for the classification of the data that was removed during the filtering process, with a certainty of classification equal to the CV-CCR. This tree also provides information about the relationships and structure between the clusters.

3. DATA

The SST data used is a monthly average of the daily SST data. The monthly average is computed using the OI v2. analysis (Reynolds, 2001). In this paper the month of choice for the cluster analysis was November. The year's chosen for this analysis range from 1940 to 2002. From these years a time series at each point in the SST image can be generated. It is these time series of the years that are by the cluster analysis to generate by the clusters, and by CART to predict the cluster membership.

4. METHODOLOGY

The first step in the analysis is to choose a hierarchical clustering scheme and the initial number of clusters to be found, done though observation of the cluster membership plot. Then an appropriate tolerance level must be selected. In this paper, the algorithm was run over eight tolerance levels 0.1, 0.5, 0.7, 0.8, 0.85, 0.9, 0.95 and 0.99. At each tolerance level the CV-CCR for that tree is computed. The best tolerance is then that which has the highest CV-CCR. From here the algorithm is run at that tolerance, and the final clusters and tree is determined. The final step is to re-classify the variables removed from the analysis by using the predictive tree model.

5. RESULTS

For this analysis, complete linkage with euclidean distance is used and the initial number of clusters is set at 20 (Figure 2).

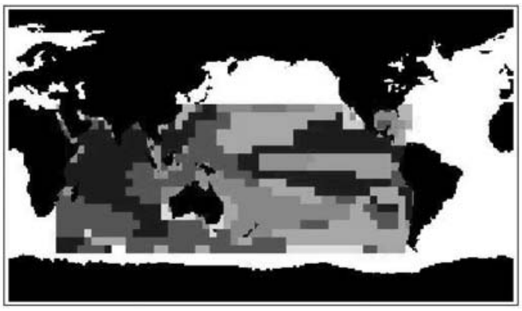

Figure 2. Original 20 clusters found by the hierarchical scheme.

The original 20 clusters in Figure 2 show that a well-defined structure exists within the SST data. Notice that most of the structure is located within the Pacific, and there are numerous smaller clusters around the edges of the image that could be removed by the filter. The tolerance for the supervised clustering scheme is now chosen by inspection of Figure 3.

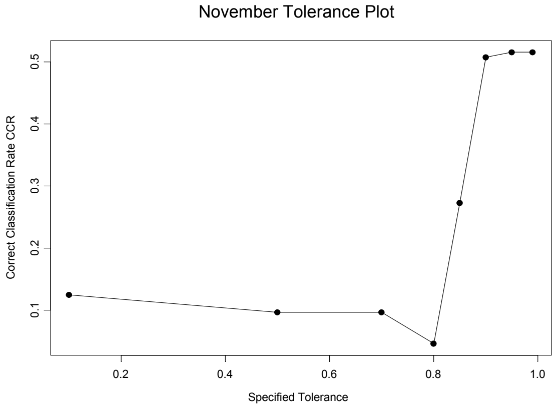

Figure 3. Specified tolerance versus CV-CCR.

Figure 3 shows that classification tolerance of 0.95, gives the clusters with the highest rate of correct classification equal to 0.50. This classification rate was achieved when the tolerance was equal to 0.90 and 0.95. As there is a significant drop in CV-CCR between 0.85 and 0.9, 0.95 was chosen as the tolerance level to ensure a stable re-classification rate.

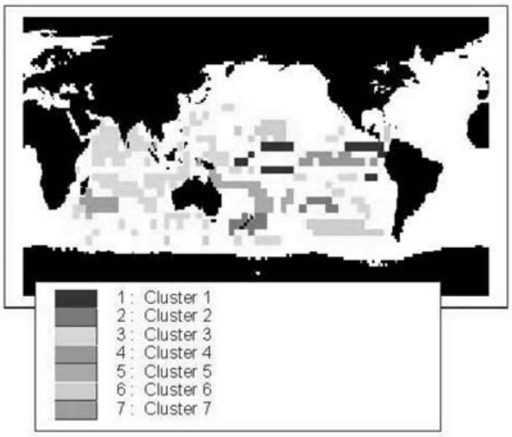

Figure 4. Final clusters selected by CART, CV-CCR = 0.50.

Figure 4, shows the clusters found after the supervised clustering algorithm had been run at a tolerance of 0.95. Overall 67 percent of the original data was removed during the supervision. The severity of the loss of data highlights the problems CART has with highly correlated large datasets. The filtered data will be re-classification using the CART model shown in Figure 5.

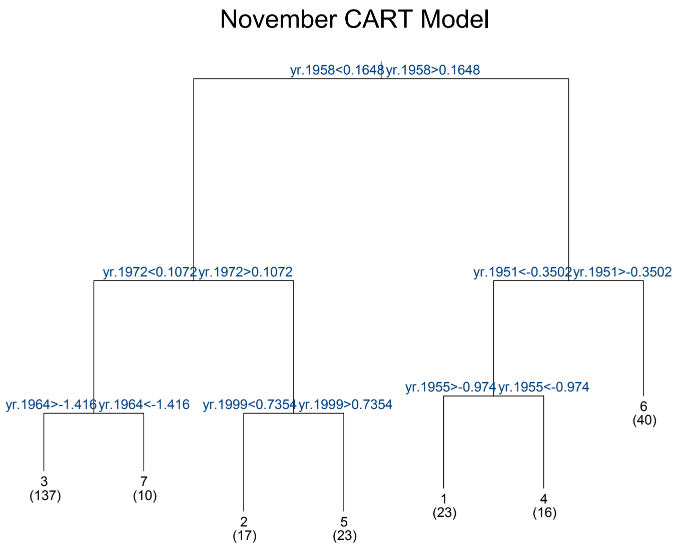

Figure 5. CART tree for the classification of the filtered clusters. (CV-CCR = 0.50)

Figure 5 is the CART model used to classify the clusters shown in Figure 4. The predictions are presented at the base of the tree. Points of interest on this tree are the years, (variables) used to classify the clusters shown in the terminal nodes. Also of interest are the sizes of the terminal nodes, presented in brackets below each prediction. These are also the size of each cluster within the model. This model is now used to re-classify the data that was removed during the supervision. The classification performance of this new data will be approximately the CV-CCR of this tree, which is 0.50. The results of this step are presented in Figure 6.

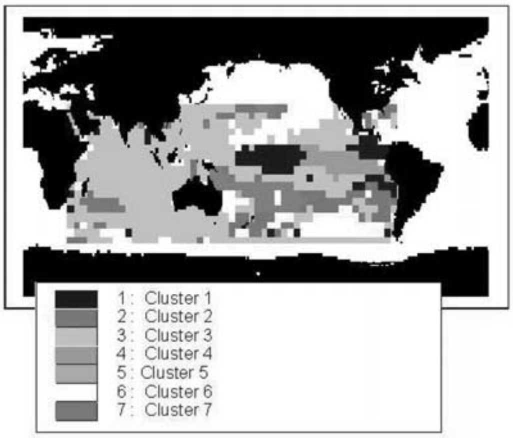

Figure 6. Tree classified clusters for all data.

Figure 6 shows the final result of the algorithm, classifying all the data. Here it can be seen that most of the structure is within the Pacific, and is similar to that found by the original hierarchical clustering scheme in Figure 2. It is noticeable that some points have been misclassified, particularly within clusters 5, 3 and 2. Also some of the clusters within equatorial Pacific have been expanded to remove the undefined clusters, this is good example of the effect of the selective classification performed by CART.

6. DISCUSSION

The trial of this supervised clustering scheme has yielded some interesting results, both statistically and climatologically. Statistically the model performed at a CV-CCR of 0.50. This with 7 unevenly sized clusters shows that there is a greater than 50% chance of classifying new data correctly. It also shows that the algorithm works, as it reduced the overfitted 20 cluster hierarchical scheme with a simpler, more useful predictive model. By observation and comparisons between the initial clustering scheme (Figure 1) and the final result of the supervision (Figure 3) it is also obvious that the clusters extracted are likely to be statistically valid.

The selective clustering by CART has generally performed well in the Pacific blurring smaller clusters together, making for simpler interpretation and more stable classification performance. However in the Indian, specifically for cluster 3 it appears to have lost some resolution around North West Australia. Cluster 3 is a particularly large cluster, with some 137 members, which is 51 percent of the correctly classified variables and 16 percent of the total number of variables. Cluster 3 also appears in both the Indian and Pacific oceans, which where shown by Drosdowsky (1993) to have statistically different signals, as they appear on different principle components. A reason for this could be that cluster 3 is clustering those variables with very little signal, and is finding a background process. Or this cluster is unable to be predicted by the CART model, and thus is poorly defined.

The tree model provides a summary of the relationships between the clusters, with respect to each other and the years that they occur. Of the years selected 4 showed a cold episode or a La Niña signal; 1972, 1951, 1999 and 1964; one showed a warm or El Niño signal; 1958 and 1951; was shown to be a neutral year. This information was extracted from the Climate Prediction Centre's (CPC) website (http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.html). From here it is possible to extrapolate and suggest that cluster 1 and cluster 4 located along the eastern equatorial Pacific which have components expressed in 1958, 1951 and 1955, also have components expressed in El Niño, La Niña and neutral years. However clusters 3, 6, 2 and 5, which are located throughout the entire region appear not to be expressed in the neutral years.

These findings agree with work previously done by Drosdowsky (1993) and Trenberth (1997), which show that regions along the equatorial Pacific are strongly related to the El Niño Southern Oscillation (ENSO) signal. To date most of the research into the ENSO signal used pre-defined regions such as the Niño 3, Niño 4 and Niño 12 regions along the equatorial pacific. This cluster analysis suggests that the strongest signals appear to between these regions and are more complicated than simple rectangles.

Like any clustering method, vastly different results could be achieved if a different set of parameters is used. This method is dependant on three inputs, the choice of the hierarchical scheme, the number of initial clusters and the specified tolerance. Of these it is feasible to test for the best specified tolerance and the best number initial number of clusters. Both these values can be extracted using an iteration approach over reasonable values. Whether the initial clustering scheme is a valid representation of reality cannot be statistically tested. As the classification performance of CART for large datasets is poor (Hastie et. al., 2001), therefore if a large dataset is supplied but an inadequate clustering scheme is specified, this algorithm then relies solely on CARTs ability to identify the clusters, and thus it is unlikely that the method will find any useful structure.

7. CONCLUSIONS

The supervised clustering method described in this paper has shown reasonable results for cluster validation in large datasets. The clusters found agree with the general regions outlined by Drosdowsky (1993). More so, the relationships between the clusters were also found through analysis of the predictive tree. With reference to the SST data this allowed for the analysis of when the signals within these clusters were strongest. This information that is useful for the determination of any climatological impact. The performance of this method is heavily dependant on the classification performance of CART for larger datasets, which can be poor (Hastie et. al., 2001). This poor performance will affect the removal of the data phase of the algorithm, resulting in large amounts of useful data being discarded. More research into a better method for the handling of the misclassified data is necessary to improve the performance of this technique.

Overall this paper has offered a simple, intuitive approach to clustering large datasets. Overcoming problems with highly correlated data and statistical cluster validation.

8. ACKNOWLEDGEMENTS

The Australian Government by way of an Australian Postgraduate Award funded the first author, Tim Hancock.

9. REFERENCES

Elizabeth J. Atkinson, Terry M. Therneau, "*An Introduction to Recursive Partitioning Using the RPART Routines"*, (2000), Mayo Foundation.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, "*Classification and Regression Trees*", (1984), Chapman and Hall, London.

Brian Everitt, "Cluster Analysis", (1974), Heinemann Educational Books, London.

G De'ath, "*New Statistical Methods for Modelling Species-Environment Relationships*", (1999), PhD Thesis, James Cook University.

Wasyl Drosdowsky, "*An Analysis of Australian seasonal Rainfall Anomalies: 1950-1987. II: Temporal Variablity and Teleconnection Patterns*", (1993), International Journal of Climatology, pp 111-149, 111.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, "*Element of Statistical Learning*", (2001), Springer, New York.

Sharon E. Nicholson, *"An Analysis of the ENSO Signal in the Tropical Atlantic and Western Indian Oceans"*, (1997), International Journal of Climatology, 17, pp. 345-375.

Kevin E Trenberth, "*The Definition of El Ni*ño", (1997), Bulletin of the American Meteorological Society, Vol 78, No 12, pp 2771-2777.

Richard W. Reynolds, Thomas M. Smith, Yan Xue, "*A New SST Climatology for the 1971-2000 Base Period and Interdecadal Changes of 30-Year SST Normal"*, (2001), Proc. 26th Annual Climate Diagnostics and Prediction Workshop.