Brooke Callan

1 December 2016

MATH 3210

Dr. Aleshunas

Course Project Report

Executive Summary

This report is about an experiment design to eventually find a rule set for classifying bridges in Pittsburgh, Pennsylvania. I will be using the Pittsburgh Bridges data set from UCI Machine Learning (Lichman, 2013) in order to run my experiment. There are 107 instances in this dataset with 12 attributes to each. There are some missing values for various attributes throughout the dataset. There are not classes assigned to each instance, but rather there are 7 attributes that predict the other five.

In order to create classes for each instance in the data set, K-means clustering will be used to determine the number of natural clusters that exist in the dataset. The missing data instances must be removed for both clustering and classification which brought the data set down to 71 instances being tested. After running K-means with 1 cluster up to 9 clusters, and examining each respective quality metric, it was determined that 4 clusters is the most appropriate for this dataset. Then, looking at the cluster vector, I was able to enter which class went with each instance into the csv file. The next step in my experiment was to perform V-fold cross validation with five different test and training groups. Each grouping of the data has 14 instances in it and then the last one has the remaining in it to be as equal as possible. Then C5.0 was ran on each of the training and test sets and create five separate classification trees. Lastly, the whole data set was used as the test and training set and a tree was created for that as well.

Extremely complex classification trees were created from each of the test/training sets as well as the whole data set. There were a total of 553, 570 rules just for simply the whole dataset tree. With such a complex tree and so many classification rules, it is almost impossible to prune the rules down to a useable classification model. Therefore, the conclusion of this project was that, with such a small data set that the algorithms could be run on, the clusters found through K-means clustering are not efficient for finding a sensible classification rule set for the Pittsburgh Bridges. The sample size of the dataset needs to be much larger in order to determine if that would help the data. Also, the variety of bridges around Pittsburgh can also be the reason for such complex classification. This is most likely why the original data set does not have assigned classes, because it was not possible to determine any without a large number.

Problem Description

In this project, I will be using K-means clustering as well as C5.0 classification methods on the Pittsburgh Bridges Data Set in order to create a classification model for how to determine what the bridge is from the data given. There are no given classes in this data set so k-means clustering will be used in order to create classes. This will then be classified with the C5.0 algorithm to determine a classification tree. Quality metrics of these classifications will be analyzed to see how appropriate the results were.

When it comes to building bridges, there are various factors engineers have to take into consideration. The first most important thing is to look at any obstacles such as rivers, valleys, and streets. Next, is to determine the dimensions of the bridges which requires them to take into consideration the use of the bridge. Then environmental factors must be analyzed to determine things like type of material and accessibility to the site. Lastly, health and safety requirements must be taken into account. (Lewis, n.d.) This all occurs before even drawing out the first sketch of the bridge.

Pittsburgh, Pennsylvania is known as the city of bridges as this city is surrounded and intertwined with three major rivers: the Allegheny River, the Ohio River, and the Monongahela River. In just one picture of downtown, over 17 bridges can be seen! The bridges all around the city play a major part in the transportation of this big city as these bridges provide access to many different parts of the city. Not every bridge is for cars and driving though. There are multiple types of bridges, colors of bridges, materials used for the bridges, and many uses of the bridges. Reports have shown that there are more than 2,000 bridges reported in the area of Pittsburgh (Popular Pittsburgh, 2015).

Analysis Technique

Three main methods will be used in this problem and they are C5.0 classification, V-fold cross validation, as well as K-means clustering. In this dataset, there is not a preclassified attribute that shows different classes, thus classification is not possible. This is where clustering comes into the picture, to create make-shift classes in order to generate classification trees. Clustering is a method where data is grouped together based on some similar attribute. There are many different types of clustering and K-means clustering will be used here.

Clustering

When starting with K-means clustering, it is important to note that every time it is run in R, it chooses a random starting point. This means that slightly different results can be found each time the same code is ran in R and this is due to how K-means works with the data. K-means uses Euclidean distance where the Euclidean distances of each instance are calculated and then the data is grouped by this. The first step in K-means is starting a randomly chosen centroid and then the Euclidean distances of each instance is calculated and this creates a first clustering. Then another centroid is calculated and then distances are calculated again where the data instances are reassigned to the new clusters. This continues until the data does not get reassigned to different clusters showing that they have reached the most appropriate clusters (K-Means Clustering, 2016).

There are two major quality metrics that are important to look at when using K-means clustering. These are total sum of squares and between sum of squares and these are important because of the use of Euclidean distances. Both quality metrics are calculated within the different clusters and the ratio of the two shows how appropriate the number of clusters are for that data set. The distance between each cluster as a whole is calculated by between sum of squares so this determines how distinct or well separated the clusters are (Kent State University). The total sum of squares is used show the compactness of each cluster, calculating the distances from the centroid to the data instances. When the ratio of these are computed, the goal is to be to have it as close one as well as level off when looking at the different number of clusters to determine what number of clusters is most appropriate. Once this is determined, these become the makeshift classes and then finding the classification rules occurs.

Classification

In order to classify the data, a classification method will be used to find rules that could identify which attribute goes with what class or cluster. Classification is a method that uses pre-classified information in order to classify the dataset. Again, there are many different classification techniques. In my project, I will be using C5.0 Classification algorithm. C5.0 Classification, this is a method in order to find a rule set on how to classify a particular dataset. C5.0 is a unique classification method that was developed off an earlier version called C4.5.

When comparing these two methods, C4.5 versus C5.0, C5.0 has been found to be have higher accuracy and less rules created than others (Pandya, 2015). This really shows how C5.0 is a better classification method and can create an accurate rule set for my experiment. A major part in creating an accurate rules set for this data set is using V-fold cross validation techniques before C5.0.

V-Fold Cross Validation

In order to make the classification trees more valid, some sort of validation must be performed. V-fold cross validation is where the dataset is divided into k number of subsets first. Then one of these subsets is used as the test set and then others are combined to be the training set. These are all ran and error matrices are computed for each iteration of the test and training sets. This allows the classification model to be trained and then tested multiple times and a more detailed look into how accurate those subsets create classification rules (Schneider, 1997). Using C5.0 with V-fold cross validation together, this breaks up the dataset and then creates classification trees that can be analyzed to create the rule set. When looking at all the trees, the most popular attributes used can be found and the other attributes will just not be considered as they were not essential to the whole data set classification.

Experiment

When looking at the Pittsburgh Bridges Data Set (Lichman, 2013), there are 107 instances with 12 different attributes. This data set does not have classes given but uniquely, there are seven specification attributes that are used to predict the other five. These seven are river, location, erected, purpose, length, lanes, and clear-g and they predict the following five attributes: through or deck, material, span, rel-l, and type. This dataset does not provide the names of the bridges where the data was collected from either. In order to prepare this dataset for the experiment, it was first transferred into an excel spreadsheet to then be converted into a csv file. Next, the as.numeric() function was used to convert any character attributes into numeric values for clustering and classification. Missing values also had to be examined to determine if they would have an impact on the results. These instances had to be removed to complete the experimentation for both clustering and classification.

My research plan begins with using k-means clustering on this data set to determine make shift classes, as they are not given in the originally data set. I will try clustering the dataset in two clusters all the way up to 9 clusters to see which one is the most appropriate. I will use the rgl package and plot 3D as well to visually see which one appears to fit the data set the best. Lastly, I will examine the quality metrics of between sum of squares and total sum of squares to see which number of clusters numerically fits the dataset the best. Once I have determine how many clusters and which entries are in each cluster, I will then perform C5.0 classification on the data set as well as 5-fold cross validation. The data will be split into test and training data sets. Each set will have 21 entries in it with the last one having 23 to ensure every entry is taken into account. I will plot the results and look at each classification tree as well as error matrices. I will then run and create a classification tree and error matrix using the entire dataset as both training and test sets to see the results. After analyzing each one, I will determine the most appropriate classification rule set by comparing results from each training and test sets as well as the last tree created. I will then present my findings of a rule set on how to classify these bridges into the makeshift classes from the most appropriate number of clusters.

Assumptions

The first assumption in my project is that all the data was correct and was entered correctly by hand from the website into a excel document, to be convert into a csv file. Next, after I compared the different quality metrics of each cluster number (1-9) for the data, I assumed that four clusters were the best fit for the data set. Other values, such as 5, could have been chosen and would be valid as well as the natural cluster number is not known. I also assumed that when the clusters were missing an instance, it was not important to the experiment and could go on without it. I also assumed that ignoring the missing value instances when clustering would also not change my results when trying to classify the dataset.

Results

After performing K-means clustering on the data, ignoring the missing data instances, it was determined that 4 clusters was the most appropriate. Comparing the quality metrics between each cluster, examining the 3D plot (Figure 1), and looking at the quality metric graph (Figure 2), four clusters was determined to be the most fitting.

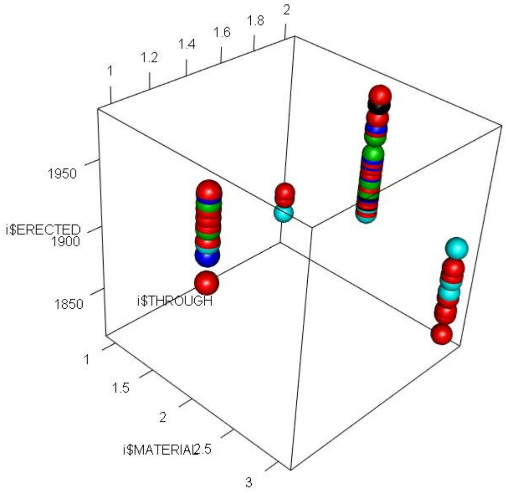

Figure 1

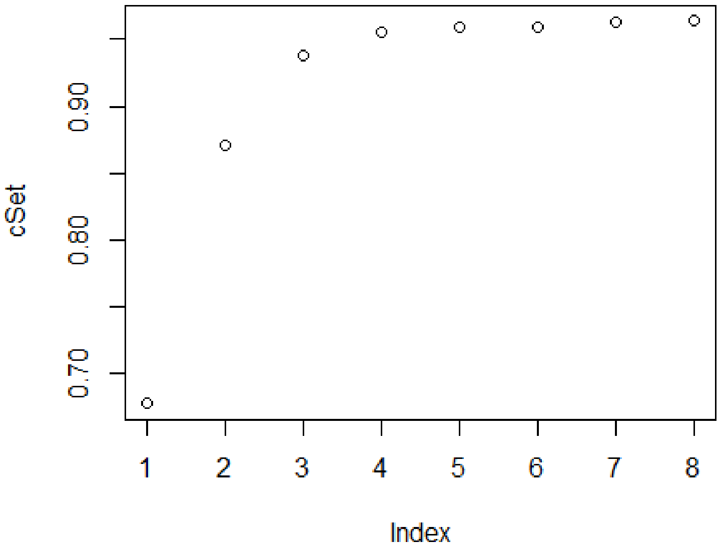

Figure 2

Then these clusters were used for classification to see if they were valid clusters and then used to create a rules set for classifying the data. The cluster labels were placed into the data set. When classifying the data, I had to remove any of the instances that weren't clustered, thus had missing data, so 71 instances were classified. These 71 instances were divided into 5 groups so each one had 14 instances and the last one had the remaining ones. Each of these were ran with their respective test and training sets in order to be able to create individual classification trees to evaluate. After performing V-fold validation and plotting those five trees as well as plotting the whole data set (Figure 3), very complex trees were revealed. By looking at these trees, it is very hard to figure out a distinct rules set for many reasons. With so many attributes, it is quite a busy tree and each one varies. If the dataset was increased with more instances, there is a better chance for finding a rules set. These trees can be used to classify the bridges but it is not the best possible model.

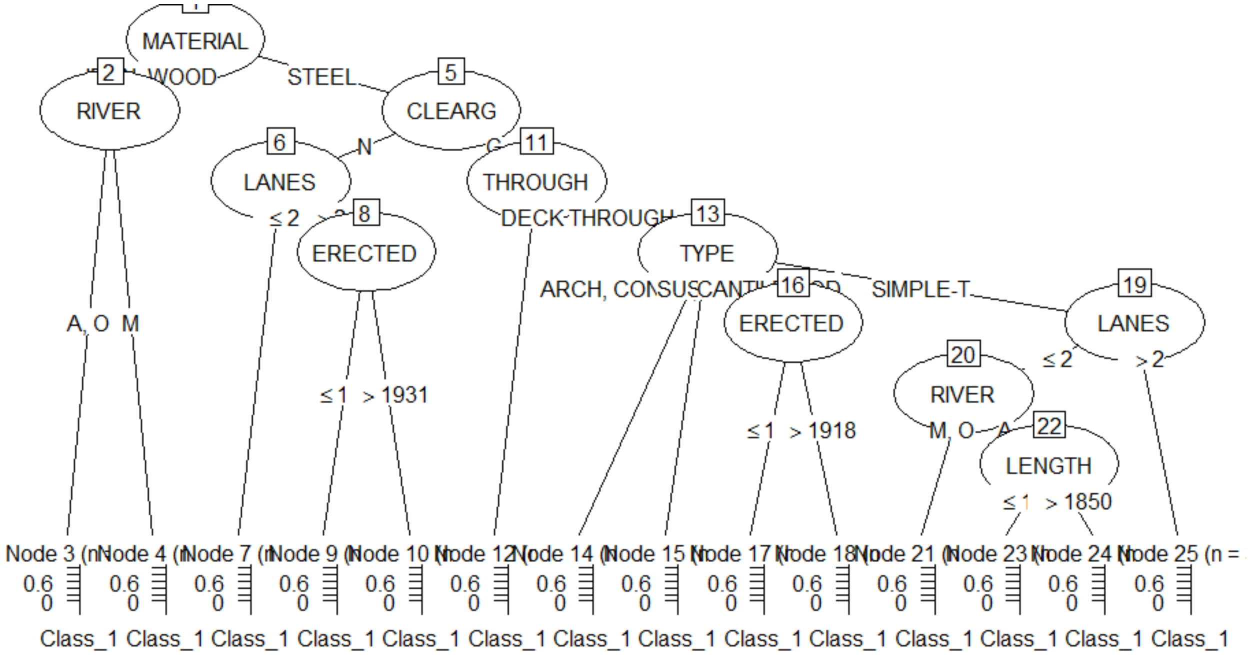

Figure 3

The final conclusion on this experiment is that the Pittsburgh Data Set does not have distinct natural clusters that can be effectively used for classifying other bridges. With a larger dataset, it could be possible to do this experiment again and see if the results change, but with 71 instances and 12 attributes, this is a very small sample. There is no distinct classification rule set for classifying Pittsburgh bridges as well as there are a total of 553,570 rules from this model.

Issues

There were many different issues that I encountered throughout this project. With having three different techniques being used in my experiment, there were many opportunities for issues to arise. First of all, when trying to cluster, K-means clustering does not work with missing data, and in my data set, there were 36 instances of 107 instances that have missing data. I had to remove these instances in order to get clusters. Also, when assigning the attributes to each cluster, there was one missing. Next, when trying to classify the data, the C5.0 algorithm continually came up with errors until it was determined that the instances that had missing classes (cluster values) had to be removed for it to run correctly. Because of this, only 71 instances could be classified. This was a sign that cautioned me to working with this data because it can be hard to get reliable results when dealing with such a small dataset size. After all of my programs ran correctly, my final issue arose as the classification trees were very complex. With over half a million rules for classifying the dataset, it is almost impossible to prune that and create a sensible rules set. The dataset of 71 instances with 12 different attributes, plus the cluster values, is very small and is a major cause of the complexity of the trees. With the amount of different bridges in Pittsburgh as well could be the cause of the complexity because when there are so many different attributes of bridges throughout the city, it can be hard to determine an actual way to classify them. More research must be done to determine if there are rules that cover most of the bridges or if there are other methods that work with missing data that would have been better. There is plenty of space to improve this experiment and possibly, this could yield different results.

References

Kent State University. (n.d.). Retrieved from Cluster Validation: http://www.cs.kent.edu/~jin/DM08/ClusterValidation.pdf

K-Means Clustering. (2016). Retrieved from A Tutorial on Clustering Algorithm: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

Lewis, J. (n.d.). *Building Model Bridges*. Retrieved from Yale-New Haven Teachers Institute : http://teachersinstitute.yale.edu/curriculum/units/2001/5/01.05.04.x.html

Lichman, M. (2013). UCI Machine Learning Repository . Irvine, CA, Universtiy of California.

Pandya, R. a. (2015). C5.0 Algoritm to Improved Decision Tree. *International Journal of Computer Applications*, Volume 117 No. 16.

Popular Pittsburgh. (2015, February 17). Retrieved from Bursting with Bridges: 2015

Schneider, J. (1997, Feburary 7). *Carnegie Mellon University*. Retrieved from School of Computer Science: https://www.cs.cmu.edu/~schneide/tut5/node42.html