C4.5 and the K-Means Clustering Algorithms

(MATH 4200 Final Project)

Clint Tomer

Executive Summary

What I have chosen to do for my final project is combine the use of the C4.5 Algorithm along with the K-Means Clustering Algorithm to classify my test set of a particular dataset. I have chosen to use the Iris.xls dataset for this experiment because it contains areas with noise and areas without noise. I will divide this dataset into two smaller sets. I will take the first set, a training set, and use it to train both K-Means and C4.5. Once K-Means has chosen the classes for the training set and C4.5 has came up with the rules for classifying the data, I will then take the second set, a test set, and run it through the rules that C4.5 established from the training set. I then have to run the whole Iris dataset through K-Means to produce the classes of the whole dataset. After I do that, I will compare the two results and see how well C4.5 classified the test set.

Here is a brief description of the dataset that I will be using. This dataset is broken down into three classes: Setosa, Versicolor, and Virginica. These three classes are determined by the following four attributes: Sepal Length, Sepal Width, Pedal Length, and Pedal Width. There are a total of 150 items in the dataset, with fifty items in each of the three classes.

One of the algorithms that I will be using, C4.5, is basically an upgrade from a similar classification algorithm, ID3. The basic idea of the ID3 algorithm is to ask questions where the answers provide the most information. The basic strategy that is used in ID3 is to choose splitting attributes with the highest information gain first. That is the amount of information associated with an attribute value that is related to the probability of occurrence (Dunham, 2003, p. 97). Once you have chosen the attributes, you will then quantify the information, which is called entropy. Entropy is used to measure the amount of uncertainty, surprise, or randomness in a dataset. The entropy will be zero if when all of the data in the set belongs to a single class (Dunham, 2003, p. 97).

The definition for entropy is:

Given the probabilities $p_1, p_2, \ldots, p_s \quad where \sum_{i=1}^{s} p_i = 1$

entropy is defined as………

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} \left(p_i \log\left(\frac{1}{p_i} \right) \right)$$

(Dunham, 2003, p. 98)

All of this information will give you the node of the tree to help classify the data. Each node contains a classifying attribute and each branch that leaves the node represents a range of values for the attribute that is going to be assigned to that node.

The algorithm C4.5 improves the ID3 algorithm in the following ways. The first improvement deals with the missing data. Whenever a decision tree is built, the missing data are ignored. This means that the gain ratio is calculated by looking at the other records that have a value for that attribute. To classify missing attribute value, the value can be predicted by the values for the other records. The second improvement deals with continuous data, which is the dividing of the data into ranges based on the attribute values for that item that are found in the training sample. The next improvement deals with the pruning of the decision tree and that is done in 1 of 2 ways. The first way is done by subtree replacement, which is where a leaf node replaces a subtree if the replacement results in an error rate close to that of the original tree. This works from the bottom of the tree to the root. The next way is done by subtree raising, which is where it replaces a subtree by its most used subtree. Raising a subtree from its current location to a node higher up in the tree does this. We also must determine the increase in error rate for this replacement. The next improvement covers issues with the rules of C4.5. The algorithm C4.5 allows classification by either decision trees or rules generated from them. Also, techniques to simplify complex rules are proposed (Dunham, 2003, p. 100). The last improvement done on C4.5 is splitting. This is done by using the largest GainRatio that ensures a larger than average information gain (Dunham, 2003, p. 101). Which is done because the GainRatio value is skewed towards splits where the size of one subset is close to that of the starting one.

The definition of GainRatio is….

$$GainRatio\ (D,S) = \frac{Gain(D,S)}{H\left(\frac{|D_1|}{|D|},\ldots,\frac{|D_s|}{|D|}\right)}$$

(Dunham, 2003, p. 101)

The second algorithm that I will be using, the K-Means algorithm is an algorithm used to cluster objects based on attributes and break them into k partitions. Its goal is to determine the k means of data generated from gaussian distributions (MacQueen (2), 1967). It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or the function:

$$V = \sum_{i=1}^{K} \sum_{j \in S_i} \left| x_j - \mu_i \right|^2$$

$$S_i, \quad i = 1,2,\ldots,K \quad and \quad \mu_i$$

Where there are k clusters is the mean point of all the points $x_j \in S_i$ (MacQueen (2), 1967)

The algorithm can be broken down into the following 4 steps (MacQueen (1), 1967):

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group means.
2. Assign each object to the group that has the closest mean.
3. When all objects have been assigned, recalculate the positions of the K means.
4. Repeat Steps 2 and 3 until the means no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The results from this experiment where very close to what I had hypothesized. The C4.5 algorithm had classed my test set of the Iris dataset with a 98.7% classification rate. It was able to classify the Setosa and Versicolor classes with a 100% (50/50) classification rate and it classified the Virginica class with a 96% (24/25) classification rate. Out of the whole test set C4.5 only miss classed 1 of the items in the set. As you can see, using C4.5 along with K-Means is a very reliable way of classifying a dataset

Problem Description

What I have chosen to do for my final project is combine the use of the C4.5 Algorithm along with the K-Means Clustering Algorithm to classify my test set of a particular dataset. I have chosen to use the Iris.xls dataset for this experiment because it contains areas with noise and areas without noise. I will divide this dataset into two smaller sets. I will take the first set, a training set, and use it to train both K-Means and C4.5. Once K-Means has chosen the classes for the training set and C4.5 has came up with the rules for classifying the data, I will then take the second set, a test set, and run it through the rules that C4.5 established from the training set. I then have to run the whole Iris dataset through K-Means to produce the classes of the whole dataset. After I do that, I will compare the two results and see how well C4.5 classified the test set.

Analysis Technique

I have chosen to use the C4.5 algorithm along with the K-Means clustering algorithm for my final project. I will take the Iris.xls dataset partition it and take my training set and run it through the K-Means clustering algorithm. The reason why I am running the data through the K-Means clustering algorithm is because it will give me my three classes of flowers that I will use to separate the data. Next, I will use the classes from the K-Means test and run that through the C4.5 algorithm to produce a decision tree. I will then use the rest of the data and run it through the decision tree I received from the C4.5 algorithm. I will do this to see how well the C4.5 algorithm classifies data.

I am going to be using the Iris.xls dataset for my experiment. This dataset is broken down into three classes: Setosa, Versicolor, and Virginica. These three classes are determined by the following four attributes: Sepal Length, Sepal Width, Pedal Length, and Pedal Width. There are a total of 150 items in the dataset, with fifty items in each of the three classes. The reason why I decided to use the Iris dataset was because it contained some “noisy” areas among two of the classes and one class can be classified without any noise. Using a dataset like Iris will allow me to test the C4.5 algorithm tolerance to noise.

C4.5 Algorithm

The C4.5 algorithm is basically an upgrade from a similar classification algorithm, ID3. The basic idea of the ID3 algorithm is to ask questions where the answers provide the most information. The basic strategy that is used in ID3 is to choose splitting attributes with the highest information gain first. That is the amount of information associated with an attribute value that is related to the probability of occurrence (Dunham, 2003, p. 97). Once you have chosen the attributes, you will then quantify the information, which is called entropy. Entropy is used to measure the amount of uncertainty, surprise, or randomness in a dataset. The entropy will be zero if when all of the data in the set belongs to a single class (Dunham, 2003, p. 97).

The definition for entropy is:

Given the probabilities $p_1, p_2, \ldots, p_s \quad where \sum_{i=1}^{s} p_i = 1$

entropy is defined as………

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} \left(p_i \log\left(\frac{1}{p_i} \right) \right)$$

(Dunham, 2003, p. 98)

All of this information will give you the node of the tree to help classify the data. Each node contains a classifying attribute and each branch that leaves the node represents a range of values for the attribute that is going to be assigned to that node.

The algorithm C4.5 improves the ID3 algorithm in the following ways. The first improvement deals with the missing data. Whenever a decision tree is built, the missing data are ignored. This means that the gain ratio is calculated by looking at the other records that have a value for that attribute. To classify missing attribute value, the value can be predicted by the values for the other records. The second improvement deals with continuous data which is the dividing of the data into ranges based on the attribute values for that item that are found in the training sample. The next improvement deals with the pruning of the decision tree and that is done in 1 of 2 ways. The first way is done by subtree replacement, which is where a leaf node replaces a subtree if the replacement results in an error rate close to that of the original tree. This works from the bottom of the tree to the root. The next way is done by subtree raising, which is where it replaces a subtree by its most used subtree. Raising a subtree from its current location to a node higher up in the tree does this. We also must determine the increase in error rate for this replacement. The next improvement covers issues with the rules of C4.5. The algorithm C4.5 allows classification by either decision trees or rules generated from them. Also, techniques to simplify complex rules are proposed (Dunham, 2003, p. 100). The last improvement done on C4.5 is splitting. This is done by using the largest GainRatio that ensures a larger than average information gain (Dunham, 2003, p. 101). Which is done because the GainRatio value is skewed towards splits where the size of one subset is close to that of the starting one.

The definition of GainRatio is….

$$GainRatio\ (D,S) = \frac{Gain(D,S)}{H\left(\frac{|D_1|}{|D|},\ldots,\frac{|D_s|}{|D|}\right)}$$

(Dunham, 2003, p. 101)

K-Means Clustering Algorithm

The K-Means algorithm is an algorithm used to cluster objects based on attributes and break them into k partitions. Its goal is to determine the k means of data generated from gaussian distributions (MacQueen (2), 1967). It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or the function:

$$V = \sum_{i=1}^{K} \sum_{j \in S_i} \left| x_j - \mu_i \right|^2$$

Where there are k clusters $S_i,\ i = 1,2,\ldots,K$ and μ_i is the mean point of all the points $x_j \in S_i$ (MacQueen (2), 1967)

The algorithm can be broken down into the following 4 steps (MacQueen (1), 1967):

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group means.

2. Assign each object to the group that has the closest mean.
3. When all objects have been assigned, recalculate the positions of the K means.
4. Repeat Steps 2 and 3 until the means no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Experiment Plan

I will first take the Iris.xls dataset and break it down. Since there are 50 items in each of the three classes, I will take 25 items from each of the three classes and use that as my partition set. This will give me a total of 75 items in each of my two sets that I will need to run this experiment. I will then take my training set and run it through the K-means clustering algorithm so it will divide my 75 items into the three classes. Before I run the data through C4.5, I must correct the items that were miss classed from the original training set. Then I will run the training set through the C4.5 algorithm to give me the rules for the decision tree. I will then take the 75 items from my test set and run them through the rules received for the training set through C4.5. Before I can see how well C4.5 classified the data, I have to run the whole Iris dataset through K-Means. Once I do this, I will be able to compare the two sets and see how well the test set was classified.

Assumptions

- The training and test sets used are a true representation of the whole dataset.
- The results are representative of the whole dataset.

Results

After comparing the results from the test set to the whole Iris dataset, I found that C4.5 had classed my test set of the Iris dataset with a 98.7% classification rate. It was able to classify the Setosa and Versicolor classes with a 100% (50/50) classification rate and it classified the Virginica class with a 96% (24/25) classification rate. Out of the whole test set C4.5 only miss classed 1 of the items in the set. As you can see, using C4.5 along with K-Means is a very reliable way of classifying a dataset.

Issues

There were only a couple of issues that I encountered whenever I did this project. I had some trouble with the naming of the folders for the algorithms. I also had some trouble getting the algorithms to produce an output file with results.

Appendices

Dunham, H. D. (2003). Data Mining: Introductory and Advance Topics. Upper Saddle River, NJ: Prentice Hall.

MacQueen, J. B. (1) (1967). "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297. Retrieved on March 28, 2006 from: http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/kmeans.html

MacQueen, J. B. (2) (1967). "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297. Retrieved on March 28, 2006 from: http://en.wikipedia.org/wiki/K-means_algorithm