C4.5 and the K-Means Clustering Algorithms

By: Clint Tomer
MATH 4200
Final Project
Outline

• Introduction to the C4.5 Algorithm
• Introduction to the K-Means Clustering Algorithm
• Dataset Overview
• Description of the Experiment
Outline cont.

- Graph of Experiment
- Hypothesized Results of Experiment
- Actual Results of the Experiment
- Experiment Conclusion
- Summary
Introduction to the C4.5 Algorithm

• An upgrade
• Basic idea is to ask questions
• Choose splitting attributes
Introduction to the C4.5 Algorithm cont.

- Entropy
  - Given probabilities $p_1, p_2, \ldots, p_s$ where $\sum_{i=1}^{s} p_i = 1$, Entropy is defined as

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} \left( p_i \log \left( \frac{1}{p_i} \right) \right)$$
Introduction to the C4.5 Algorithm cont.

• C4.5 improves ID3 in the following ways:
  – Missing Data
  – Continuous Data
  – Pruning
    • Subtree Replacement
    • Subtree Raising
Introduction to the C4.5 Algorithm cont.

- C4.5 improves ID3 in the following ways cont.
  - Rules
  - Splitting
    - GainRatio

\[
GainRatio \ (D, S) = \frac{Gain(D, S)}{H\left(\frac{|D_1|}{|D|}, \ldots, \frac{|D_s|}{|D|}\right)}
\]
Introduction to the K-Means Clustering Algorithm

- Cluster objects
- Determine the K-Means
- Objects attributes form a vector space
Introduction to the K-Means Clustering Algorithm cont.

- The objective K-Means tries to achieve is to minimize total intra-cluster variance, or the function 

\[
V = \sum_{i=1}^{K} \sum_{j \in S_i} (x_j - \mu_i)^2
\]

where there are k clusters \( S_i, \ i = 1, 2, \ldots, K \) and \( \mu_i \) is the mean of all points \( x_j \in S_i \).
Introduction to the K-Means Clustering Algorithm cont.

- The K-Means Clustering Algorithm can be broken down into the following steps
  1. Place k points into the space represented by the objects that are being clustered.
  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.
Dataset Overview

• Iris.xls dataset
• 3 Classes
  – Setosa
  – Versicolor
  – Virginica
• 4 Attributes
  – Sepal Length
  – Sepal Width
  – Pedal Length
  – Pedal Width
• 150 Items
  – 50 of each class
Description of Experiment

• K-Means
• C4.5
• C4.5 Rules
• K-Means
• Compare
• Classification
Graph of the Experiment

Iris → K-Means → Classes → C4.5 → Rules

Te Tr → Tr

Compare

Results Tr → Results Te
Hypothesized Results

- Very accurate
- Close to 100% classification rate
Results of the Experiment

- Setosa
- Versicolor
- Virginica
- Classification of data
Experiment Conclusion

- C4.5
- K-Means
- Classification
Summary

• C4.5 Algorithm
• K-Means Clustering Algorithm
• Dataset
• Experiment