Project 4

Jake Carlson

November 24, 2017

Abstract

Clustering

Contents

Attribute	Scale	Values
Age	Interval	17, 22, 27, 32, 37, 42, 47, 52, 57, 62, 67, 72, 75
Education	Ratio	1 - 22
LOS	Interval	1, 3, 7, 12, 17, 22, 27, 32, 35
Pay	Ratio	1 - 401,000
SupervisoryStatus	Nominal	0, 1

Table 1: Final Data Set Attributes

1 Business Understanding

This report will focus on clustering analysis of the federal payroll data obtained by BuzzFeed News through the Freedom of Information Act. Specifically, k-means and hierarchical clustering will be used to find groups of emplyees with similar attributes in the federal government.

Armed with an undestanding of what groups exist within each agency, agency leaders can work to create employee teams that are balanced with respect to the features each group holds. We can also see what features create the largest separation between the subgroups.

2 Data Preparation

I will prepare my data in a similar fashion to Project 3 [3]. I will take the middle of the age range for the age value of each employee. I will do the same thing for length of service. Education will be converted to an integer representing the number of years needed to achieve the degree the employee holds so that all education values are ratio scaled. I will also make supervisory status binary where a one indicates an employee is a supervisor and a zero indicates an employee is not a supervisor. The final list of attributes to be used for clustering is given in Table 1.

All of these attributes are then scaled so that the values are between -1 and 1. With all of these attributes scaled, I can use Euclidean distance as my distance metric for clustering because all of the attributes are on the same scale.

3 Modeling

4 Evaluation

References

[1] Jake Carlson *CSE 5331 - Data Mining Project 1* `https://github.com/jakecarlson1/data-mining-projects/blob/master/project-1/report/carlson-project-1.pdf`

[2] Jake Carlson *CSE 5331 - Data Mining Project 2* https://github.com/jakecarlson1/data-mining-projects/blob/master/project-2/report/carlson-project-2.pdf

[3] Jake Carlson *CSE 5331 - Data Mining Project 3* https://github.com/jakecarlson1/data-mining-projects/blob/master/project-3/report/carlson-project-3.pdf