Project 3

Jake Carlson

November 4, 2017

Abstract

This report examines federal payroll data in the years 2005 and 2013. The data will be transformed into transaction data so that association rule mining can be performed. By looking at the differences in maximal, closed, and frequent itemsets, I will be able to understand common features about employees under the presidencies of George W. Bush and Barack Obama.

Contents

1 Business Understanding

In this report I will again be examining the federal payroll data obtained by BuzzFeed News through the Freedom of Information Act. I will look at the years 2005 and 2013 so I can gauge how the government changed as the presidency transitioned from George W. Bush to Barack Obama. This report will focus on association rule mining. The payroll data will be formatted as transaction data, where each employee is treated as a bag of features where all attributes that apply to them are placed in their bag. With the data formatted in this way, I can use the arules package to find items that occur together. These items can be treated as rules, where the occurance of some items tend to imply the occurance of another item. With these rules, along with statistics about the rule such as the support, or the ratio of instances that express the rule, and confidence, or the conditional probability that the predicted item occurs in an instance which contains all remaining items for the rule, I will build a picture of the key differences between groups of federal employees under the two presidents.

2 Data Preparation

I will start with the data as I prepared it for classification in Project 2. This data has been cleaned such that all unknown values were replaced with NA, Age and Length of Service were adjusted to take the middle year for the range of years given in the raw data, and Pay has ben discretized into the pay ranges given in Table 1. In addition, Education is descritized into the groups given in Table 2.

I will get rid of attributes that I am not interested in for association rule mining. I will drop the Agency attribute because AgencyName already encodes this data. Likewise, I will drop Station because region records the state an employee works in.

All numeric fields will be discretized based on frequency **explain why**. SupervisoryStatus must be treated as a factor so that general employees are not grouped with supervisors. The final list of attributes is given in Table 3.

These attributes are then transformed into a transaction item matrix. The most frequent items show us what attributes are the most common for employees to have. Summaries for both the 2005 and 2013 transaction data sets are given in Table 4. We see that for both years, the most common employees are non-supervisors. The majority of employees are also college educated and have been working in the federal government for less than twelve years. Most employees are younger than 47 years old and the most frequent employment category is Administrative.

Pay Ranges
<30k
30-50k
50-70k
70-90k
90-110k
>110k

Table 1: The Pay Ranges Used To Descretize Pay

Group	Education Levels	Description
Elm	0, 1	Reached or completed elementary school
HS	3, 4, 5, 6	Reached or completed high school or an occupational program
Col	7, 8, 9, 10, 11, 12, 13	Reached or completed college with a Bachelor's degree
Grad	14, 15, 16, 17, 18, 19, 20	Any level of graduate studies, excluding a Doctorate
Doc	21, 22	A Doctorate or Post-Doctorate degree

Table 2: Ordinal Education Groups

Attribute	Scale	Range
AgencyName	Nominal	The name of each agency
region	Nominal	The name of the state
Age	Interval	[17,47), [47,57), [57,75]
Education	Ordinal	Elm, HS, Col, Grad, Doc
LOS	Interval	[1,22), 22, [27,35]
Category	Nominal	P, A, T, C, O, B
Pay	Ordinal	<30k, 30-50k, 50-70k, 70-90k, 90-110k, >110k
SupervisoryStatus	Nominal	2, 4, 5, 6, 7, 8

Table 3: Final Data Set Attributes

2005	Elements	4,720,680
	Item	Count
	SupervisoryStatus=8	4,081,468
	Education=Col	2,190,442
	Age=[17,47)	2,087,004
	LOS=[1,12)	1,900,001
	Category=A	1,716,162
	Itemset Length	Count
	4	18
	5	2529
	6	85,052
	7	2,034,281
	8	2,598,800
2013	Elements	5,323,899
	Item	Count
	SupervisoryStatus=8	4,502,511
	Education=Col	2,574,899
	LOS=[1,12)	2,568,556
	Age=[17,47)	2,276,609
	Category=A	2,034,255
	Itemset Length	Count
	5	126
	6	13,770
	7	2,173,306
	8	3,136,697

Table 4: Summaries of 2005 and 2013 Transaction Data Sets

3 Modeling

4 Evaluation