CIS 635 Data Mining

Project

Description

The project this semester consists of building a pipeline for periodic analysis of (generated) employee medical data. You will need to prepare the data by combining two data sources, cleaning the data and preparing it for processing. You will also need to choose the best classification technique. The end result is that you will create a pipeline to prepare and analyze the data and write an executive summary report of your work.

Data

The files that you are given contain data for 6,781 employees. You will get weekly training and test data from 2 sources. All files have an id for each employee which uniquely identifies the employee. Thus employee 3473 in the A file is the same employee as 3473 in the B file. The problem is that the id numbers are not in order. Also, there are some missing values that need to be handled.

Source A files contain some vital measurements taken weekly:

variable	type	description	acceptable values
id	unique key		
temp	numeric	patient's temperature	90 - 106
bpSys	numeric	blood pressure (systolic)	90 - 150
vo2	numeric	VO^2 max	10 - 70
throat	numeric	throat culture	80 - 120
atRisk	factor	virus test	0=no virus, 1=has virus

Source B files contain the results of weekly surveys:

variable		description	acceptable values
id	unique key		
headA	factor		0 to 9
bodyA	factor		0 to 9
cough	factor		0=no, 1=yes
runny	factor		0=no, 1=yes
nausea	factor		0=no, 1=yes
diarrhea	factor		0=no, 1=yes

Procedure

Follow the steps below to process the data files:

1. preprocess the data
   a) join the data sources using the unique key
   b) locate and fix missing data (you can choose to delete the row, the column or estimate the value of the missing data)
   c) locate and fix noise (same choices as for missing data)
   d) count the outliers for each attribute (do not change their values)
   e) calculate the bin ranges for discretization of the attribute throat (will not be used later)
2. cluster the data to learn more about it.
   a) normalize data using minMax
   b) use kmeans to find clusters
   c) display the centroids using the original scale (not minMax averages)
   d) describe clusters in a meaningful way
   e) find the optimum number of clusters using the “knee” method of plotting SSE
3. run tests to compare classifiers and make a choice based on performance
   a) use decision tree (rpart), naive Bayes, K nearest neighbor, SVM (linear or radial basis kernel) and ANN (use 5 hidden nodes).
   b) choose the best model (algorithm) and defend your choice using statistics, clustering results, scatter plots, boxplots and/or histograms.
4. write an R script that can be run weekly to
   a) clean and merge the data sources (two files for training and two for test)
   b) build classification model with the training data
   c) predict atRisk for employees using the test data and write predictions to a text file
5. write a one paragraph (200 words maximum)
   a) describe the nature of the data
   b) describe the weekly prediction (including the quality of the prediction using precision, recall and accuracy)

Hand in

- handin sheet with:
  - with the values filled in. Change your answers to **bold** so that they stand out.
  - the weekly R script
  - executive summary

Details

Data (important – read this):

The data set you are given consists of 4 files, the source A and B data for both training and test. You will notice that there are 10 sets (10 sets of 4 files, 40 files total). You are assigned a number from 0 to 9. That number will determine the data set you are to use. You can find your data set number from the table on the next page. It is important that you use the assigned data set. If you do not, you will not receive credit for this project. Also, as always, if there is any evidence of cheating you will not receive credit for this report.

Rubric :

description	points
cleaning	
merging	2
outliers	2
noise	2
discretization	3
clustering	
choice of k	5
centroids	3
cluster descriptions	4
classification	
results	3
choice and defense	10
R script	12
executive summary	4
total	50

Data set assignments:

Section 01 (Wed night)			Section 02 (Mon night)	
G Number	data set nbr		G Number	data set nbr
G00244974	1		G01057754	2
G00892560	4		G01094706	0
G01083302	4		G01665722	1
G01289302	9		G02278762	8
G01480185	4		G02333188	2
G01690040	8		G02341698	4
G02336309	2		G02345343	5
G02344955	5		G02345383	9
G02360629	2		G02381692	2
G02381702	0		G02385419	6
G02383772	5		G02406497	4
G02385327	6		G02449118	3
G02385674	7		G02484927	5
G02406870	3		G02485542	1
G02406980	7		G02486543	0
G02448581	9		G02487344	7
G02448771	6		G02492669	7
G02448982	1			
G02449520	3			
G02471660	0			
G02483873	8			
G02484205	9			
G02484905	5			
G02485361	8			
G02486757	1			