Exploring RapidMiner

K-means Clustering

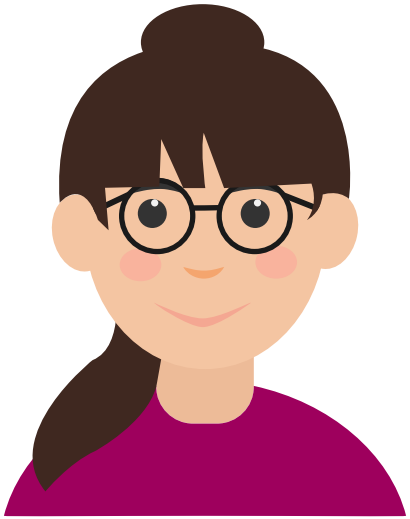
PL10

[Hugo Peixoto] – 2018/2019, Universidade do Minho

Material

http://hpeixoto.github.io/dc

CONTEXT AND PERSPECTIVE



Sonia is a program director for a major health insurance provider.

Recently she has been reading in medical journals and other articles, and found a strong emphasis on the influence of weight, gender and cholesterol on the development of coronary heart disease.

She begins brainstorming ideas for her company to offer weight and cholesterol management programs to individuals who receive health insurance through her employer.

As she considers where her efforts might be most effective, she finds herself wondering if there are natural groups of individuals who are most at risk for high weight and high cholesterol, and if there are such groups, where the natural dividing lines between the groups occur.

BUSINESS UNDERSTANDING

Sonia's goal is to identify and then try to reach out to individuals insured by her employer who are at high risk for coronary heart disease because of their weight and/or high cholesterol.

She understands that those at low risk, that is, those with low weight and cholesterol, are unlikely to participate in the programs she will offer.

She also understands that there are probably policy holders with high weight and low cholesterol, those with high weight and high cholesterol, and those with low weight and high cholesterol.

She further recognizes there are likely to be a lot of people somewhere in between.

DATA UNDERSTANDING

Using the insurance company's claims database, Sonia extracts three attributes for 547 randomly selected individuals.

The three attributes are the insured's **weight** in pounds as recorded on the person's most recent medical examination, their last **cholesterol** level determined by blood work in their doctor's lab, and their **gender**.

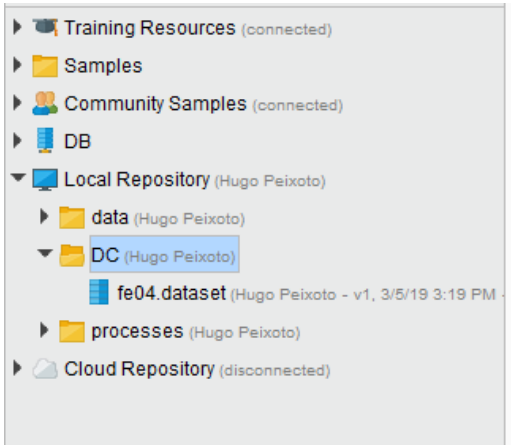
As is typical in many data sets, the gender attribute uses 0 to indicate Female and 1 to indicate Male.

We should remember as we do this that means are particularly susceptible to undue influence by extreme outliers, **so watching for inconsistent data** when using the k-Means clustering data mining methodology is very important.

DATA PREPARATION

Download csv: http://hpeixoto.github.io/dc/pl10/pl10.dataset.csv

Import csv to rapidminer repository.



Check results tab and inspect metadata view of the imported csv.

MODELING

The 'k' in k-means clustering stands for some number of groups, or clusters. The aim of this data mining methodology is to look at each observation's individual attribute values and compare them to the means, or in other words averages, of potential groups of other observations in order to find natural groups that are similar to one another.

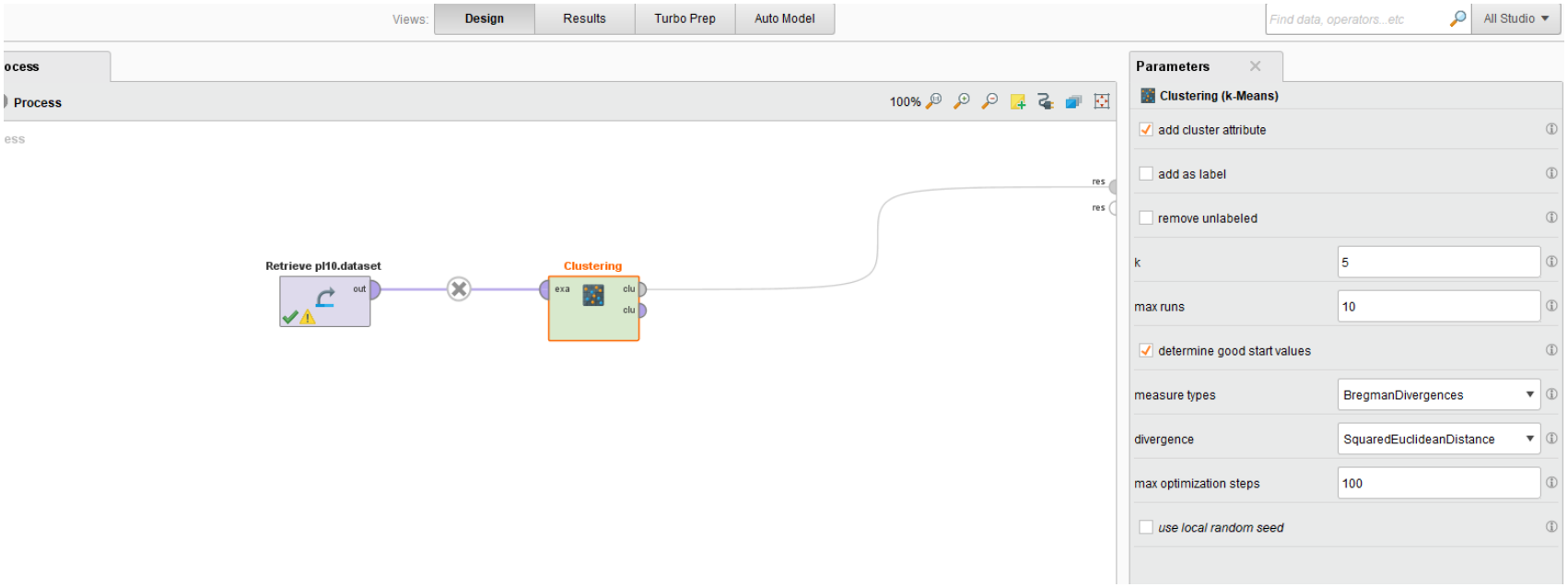
The k-means algorithm accomplishes this by sampling some set of observations in the data set, calculating the averages, or means, for each attribute for the observations in that sample, and then comparing the other attributes in the data set to that sample's means.

The system does this repetitively in order to 'circle-in' on the best matches and then to formulate groups of observations which become the clusters. As the means calculated become more and more similar, clusters are formed, and each observation whose attributes values are most like the means of a cluster become members of that cluster.

MODELING

On the Operators tab in the lower left hand corner, use the search box and begin typing in the word *clustering*.

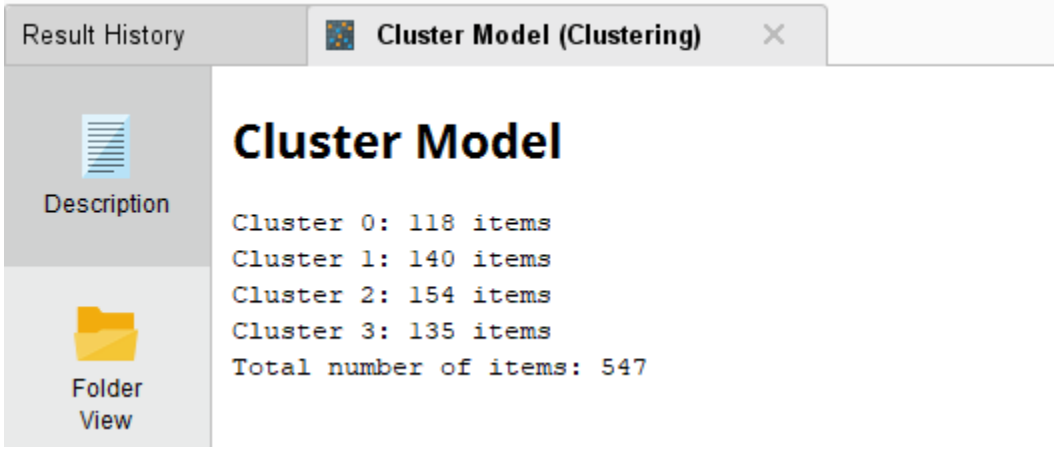
The tool we are looking for is called *k-means*. Drag and drop and click *Run*.



MODELING

But as discussed in the Organizational Understanding section earlier in the chapter, Sonia has already recognized that there are likely a number of different types of groups to be considered. Simply splitting the data set into two clusters is probably not going to give Sonia the level of detail she seeks. Because Sonia felt that there were probably at least 4 potentially different groups, let's change the k value to four.

The distribution of observations across our four clusters.



MODELING

We could go back at this point and adjust our number of clusters, our number of 'max runs', or even experiment with the other parameters offered by the k-Means operator

EVALUATION

Recall that Sonia's major objective in the hypothetical scenario posed at the beginning of the chapter was to try to find natural breaks between different types of heart disease risk groups.

Centroid Table

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Weight	152.093	106.850	184.318	127.726
Cholesterol	185.907	119.536	218.916	154.385
Gender	0.441	0.543	0.591	0.459

gives the means for each attribute in each of the four clusters we created

EVALUATION

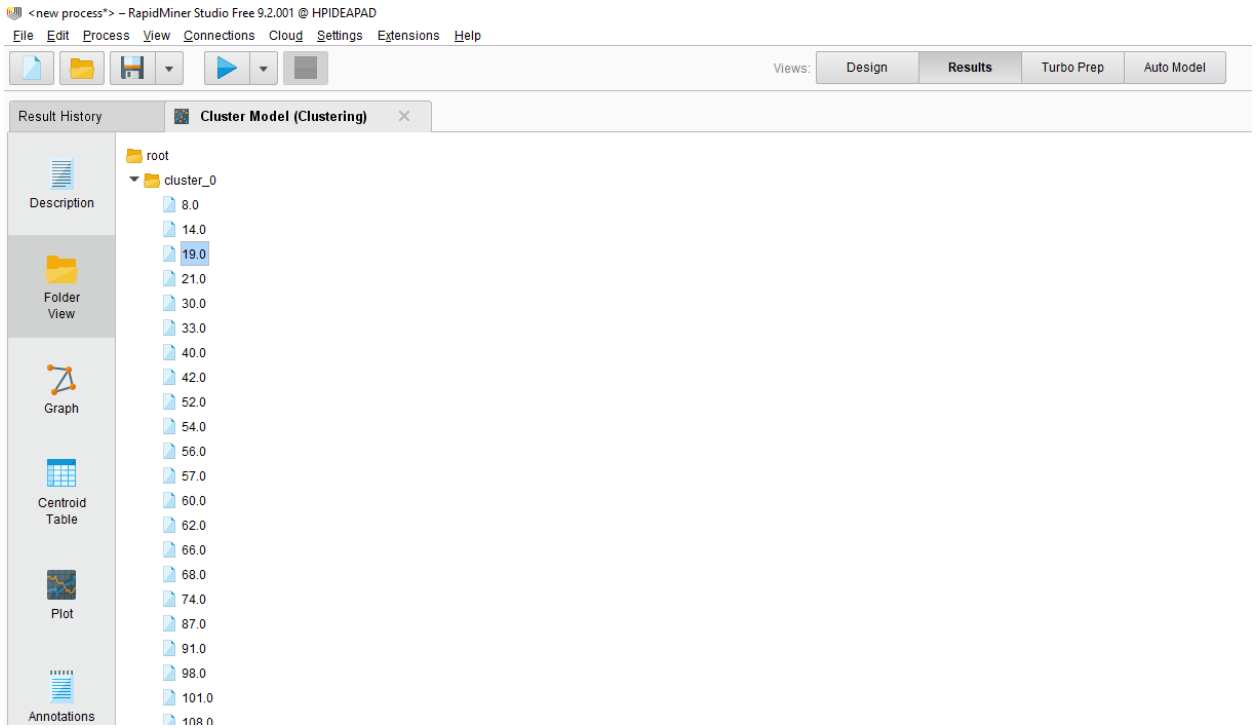
We see in this view that cluster 0 has the highest average weight and cholesterol. With 0 representing Female and 1 representing Male, a mean of 0.591 indicates that we have more men than women represented in this cluster.

Knowing that high cholesterol and weight are two key indicators of heart disease risk that policy holders can do something about, Sonia would likely want to start with the members of cluster 0 when promoting her new programs.

So we know that cluster 0 is where Sonia will likely focus her early efforts, but how does she know who to try to contact?

EVALUATION

Folder View

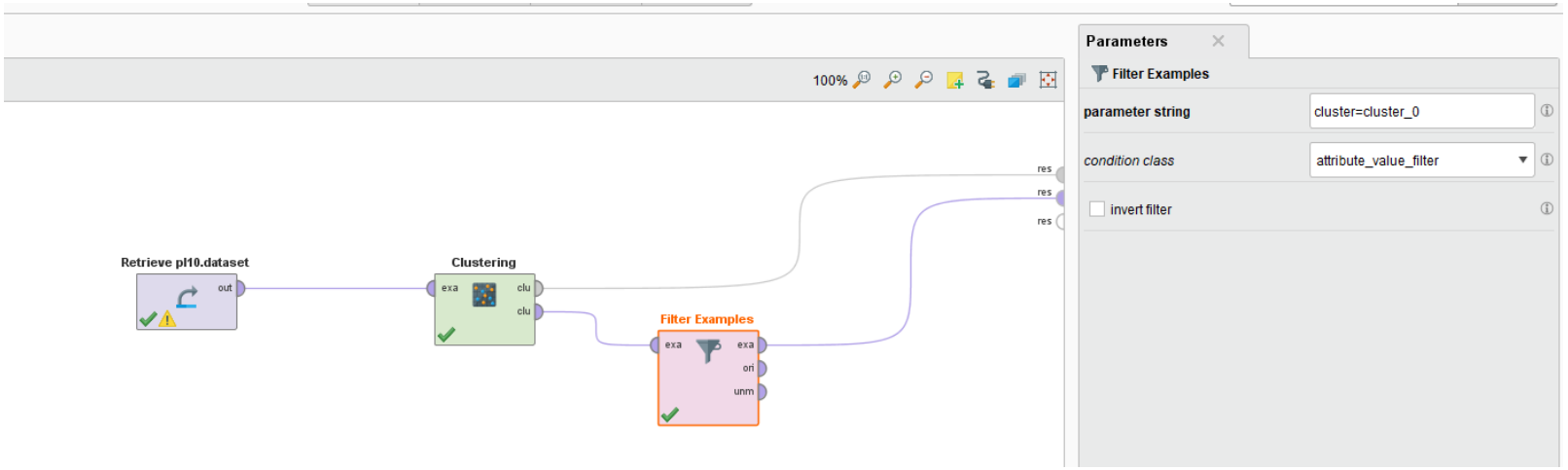


EVALUATION

The means for cluster 0 were just over 184 pounds for weight and just under 219 for cholesterol. The person represented in observation 6 is heavier and has higher cholesterol than the average for this highest risk group. Thus, this is a person Sonia is really hoping to help with her outreach program.

DEPLOYMENT

Using the search field in the Operators tab, locate the Filter Examples operator and connect it to your k-Means Clustering operator. Note that we have not disconnected the clu (cluster) port from the 'res' (result set) port, but rather, we have connected a second clu port to our exa port on the Filter Examples operator, and connected the exa port from Filter Examples to its own res port.



DEPLOYMENT

With the high risk group having weights between 167 and 203 pounds, and cholesterol levels between 204 and 235 (these are taken from the Range statistics in Figure 6-10), she could return to her
company's database and issue a SQL query like this one:

```
SELECT First_Name, Last_Name, Policy_Num, Address, Phone_Num
FROM PolicyHolders_view
WHERE Weight >= 167
AND Cholesterol >= 204;
```

Try cluster=cluster_1 and build the SQL question.

SUMMARY

k-Means clustering is a data mining model that falls primarily on the side of Classification.

For this example, it does not necessarily predict which insurance policy holders will or will not develop heart disease. It simply takes known indicators from the attributes in a data set, and groups them together based on those attributes' similarity to group averages.

Because any attributes that can be quantified can also have means calculated, k-means clustering provides an effective way of grouping observations together based on what is typical or normal for that group.

k-Means clustering is very **flexible** in its ability to group observations together. The k-Means operator in RapidMiner allows data miners to set the number of clusters they wish to generate, to dictate the number of sample means used to determine the clusters, and to use a number of different algorithms to evaluate means. **While fairly simple in its set-up and definition, k-Means clustering is a powerful method for finding natural groups of observations in a data set.**

FE08

Exploring RapidMiner

K-means Clustering

PL10

[Hugo Peixoto] – 2018/2019, Universidade do Minho