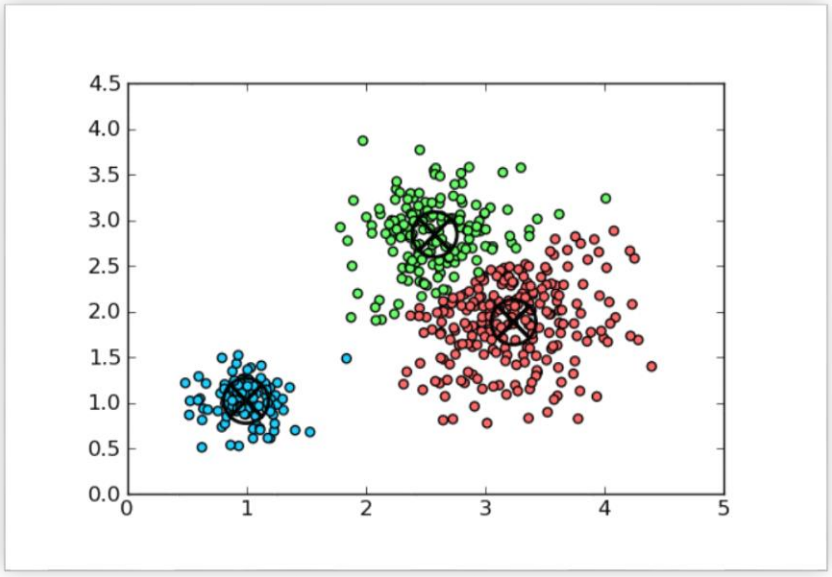


K-means Clustering

USING RAPIDMINER

Sanchit Kumar | Data Warehousing and Data Mining | April 27, 2016

Problem Statement

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. The research she's read confirms time after time that there is a connection between these three variables, and while there is little that can be done about one's gender, there are certainly life choices that can be made to alter one's cholesterol and weight. 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.

Algorithm Used

K-MEANS ALGORITHM

Formally, given a data set, D, of n objects, and k, the number of clusters to form, the k-means algorithm organizes the objects into k partitions, where each partition represents a cluster. The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters in terms of the data set attributes.

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Data Set Used

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 will use this sample data from Sonia's employer's database to build a cluster model to help Sonia understand how her company's clients, the health insurance policy holders, appear to group together on the basis of their weights, genders and cholesterol levels.

A data set has been prepared for this example, and is available as Chapter06DataSet.csv on the book's (Data Mining for the Masses) companion web site.

Applications of the Algorithm

K-means clustering is very flexible in its ability to group observations together. 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. It also helps us understand where one group begins and the other ends, or in other words, where the natural breaks occur between groups in a data set.

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.

Screenshots

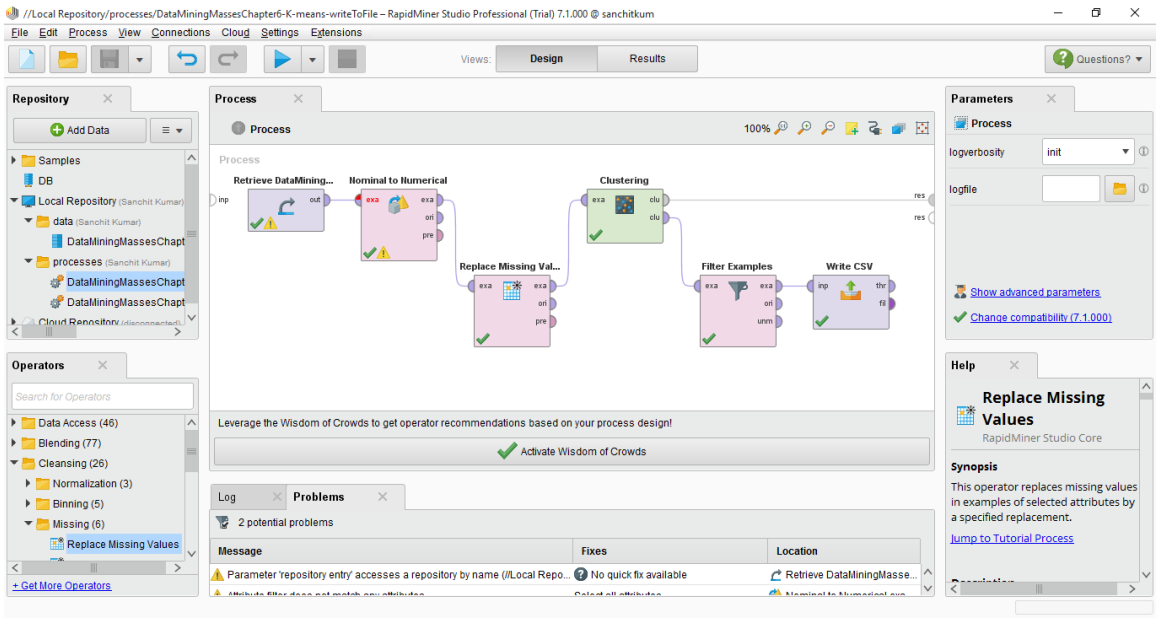


Fig 1. Process View

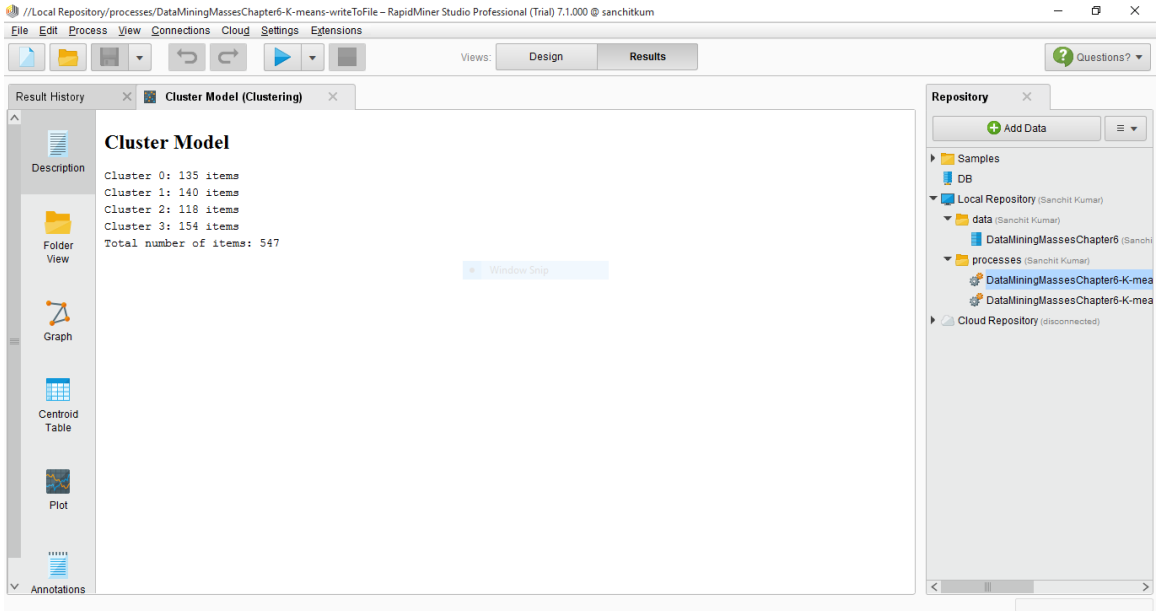


Fig 2. Cluster Model

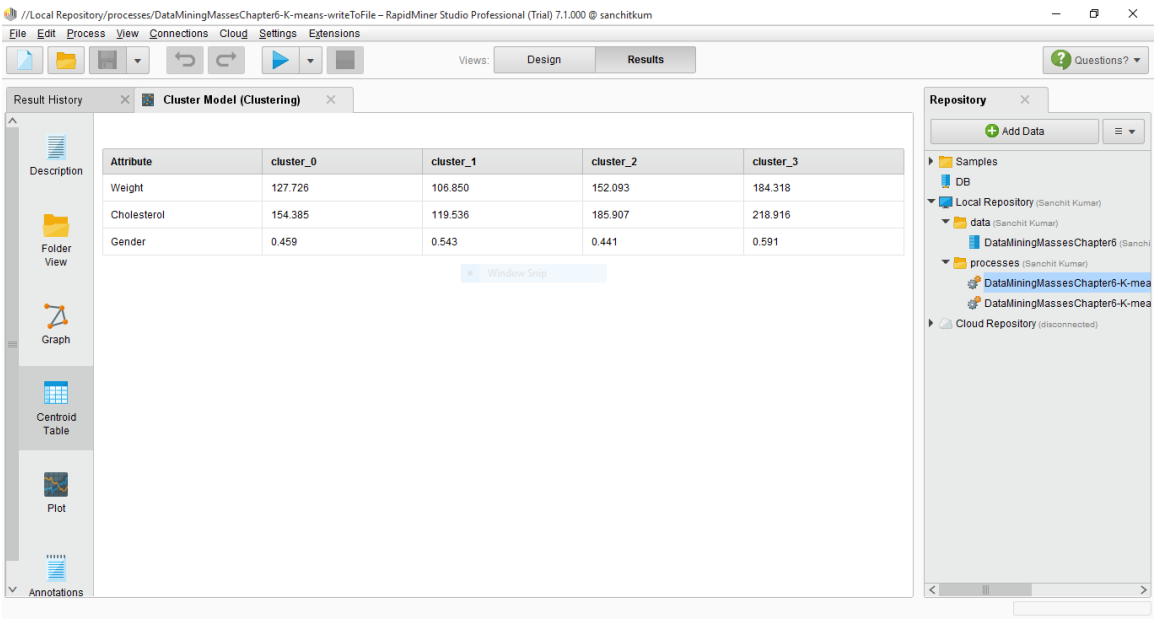


Fig 3. Centroid Table

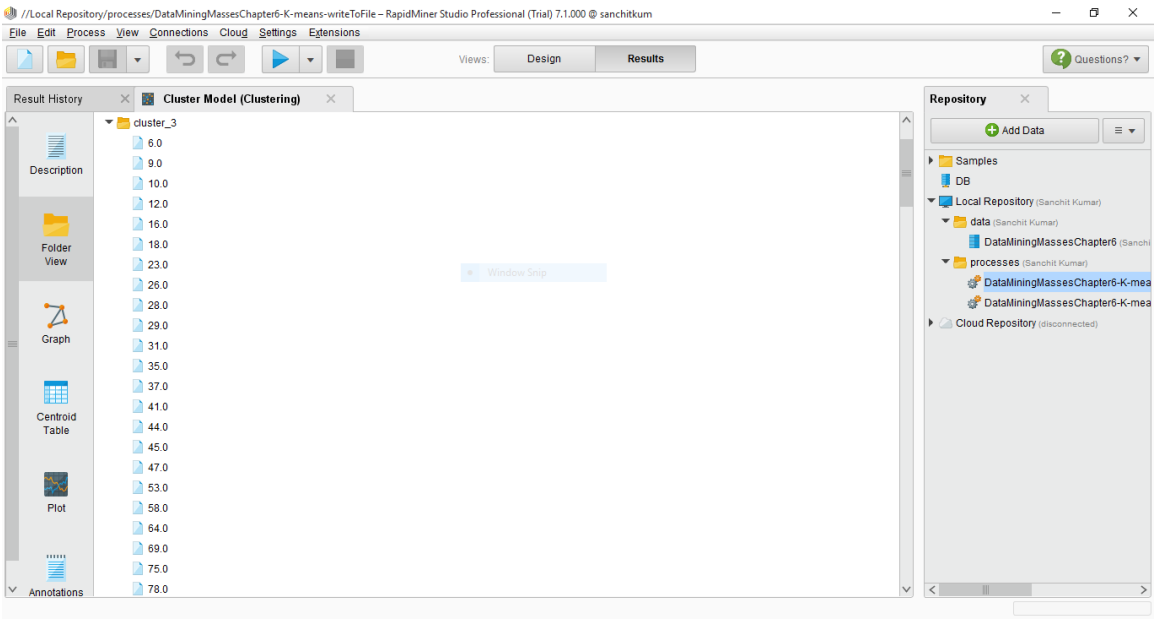
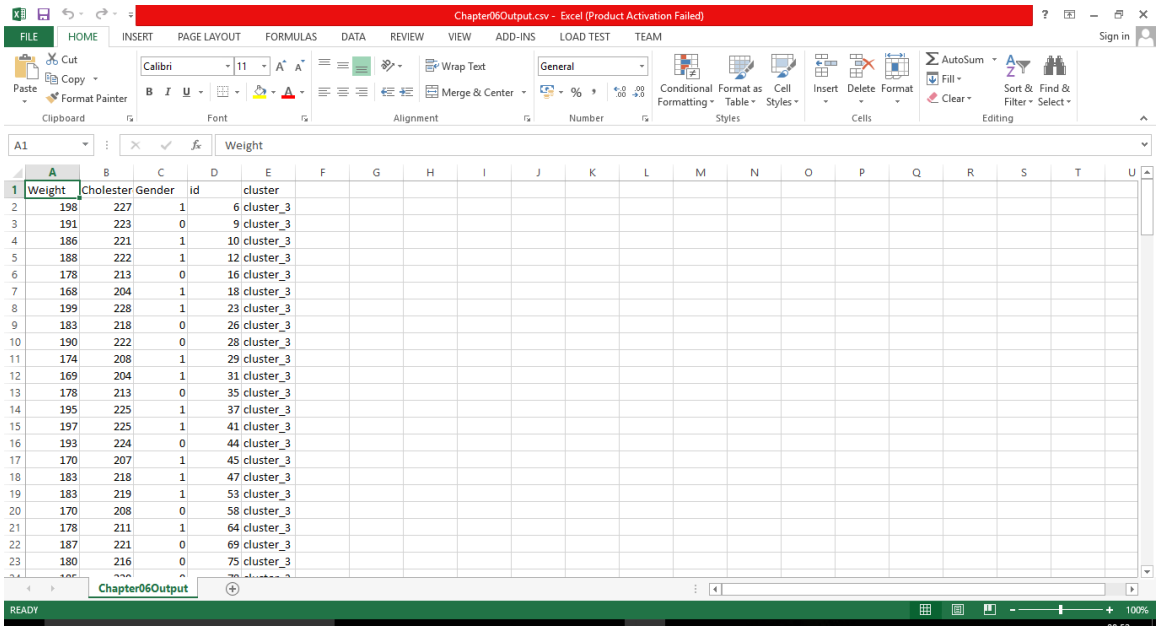


Fig 4. Folder View of Cluster 3



	A	B	C	D	E
1	Weight	Cholester	Gender	id	cluster
2	198	227	1	6	cluster_3
3	191	223	0	9	cluster_3
4	186	221	1	10	cluster_3
5	188	222	1	12	cluster_3
6	178	213	0	16	cluster_3
7	168	204	1	18	cluster_3
8	199	228	1	23	cluster_3
9	183	218	0	26	cluster_3
10	190	222	0	28	cluster_3
11	174	208	1	29	cluster_3
12	169	204	1	31	cluster_3
13	178	213	0	35	cluster_3
14	195	225	1	37	cluster_3
15	197	225	1	41	cluster_3
16	193	224	0	44	cluster_3
17	170	207	1	45	cluster_3
18	183	218	1	47	cluster_3
19	183	219	1	53	cluster_3
20	170	208	0	58	cluster_3
21	178	211	1	64	cluster_3
22	187	221	0	69	cluster_3
23	180	216	0	75	cluster_3

Fig. 5. Filtered View of the data belonging to Cluster 3

Evaluation

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. Using the k-Means operator in RapidMiner, we have identified four clusters for Sonia, and we can now evaluate their usefulness in addressing Sonia's question.

We see in the screenshots that cluster 3 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 3 when promoting her new programs. She could then extend her programming to include the people in clusters 2 and 0, which have the next incrementally lower means for these two key risk factor attributes.

References

1. Book: Data Mining for the Masses - Dr. Matthew A North
2. Book: Data Mining: Concepts and Techniques - Han, Kamber, Pei
3. Dataset: Data Mining for the Masses - Site
4. Video: K-Means Clustering in RapidMiner - YouTube