Assignment

Name: Yuqiao Chen

Student Number: 29836354

Module Title: Data Analytics and Mining

Module Code: CSMDM16 / CSMDM21

Problem #1

Task1:

First, we load the data into the Knime by using the “File Reader” node. The path of “teeth.csv” in my laptop is “C:\Users\CYQ\knime-workspace\assignment\data\teeth.csv” 表格

描述已自动生成

Next we use “Pie Chart” to see the number of records. Here we can the pie chart shows **30** records

图形用户界面, 文本, 应用程序

描述已自动生成图片包含 图表

描述已自动生成

Then we use “Statistics” node to statistics on attributes.

图形用户界面, 应用程序

描述已自动生成

In the image below, we can clearly see that there are **8** attributes in total. The chart details that all 8 attributes are integer values, and we can visually get the **mean value** of each attribute.

表格

中度可信度描述已自动生成

We calculate the **range** of each attribute by using “Column Filter” and “Math Formula”. Then we can get 8 attribute range. For example, we can get the range for “TopInc” as 3.

图形用户界面, 应用程序, Word

描述已自动生成图形用户界面, 应用程序

描述已自动生成表格

描述已自动生成

Then we use “Histogram(local)” node to print all histogram for each attribute, I use “TopInc” as an example and other attributes are same as this one. Config the “Histogram(local)” node by changing the “Binning column” and “aggregation columns” to “TopInc”, then execute the node, we can get the histogram of “TopInc” attribute.

In the following picture we can clearly see the number of different values of "TopInc".

图形用户界面, 应用程序

描述已自动生成 图表, 条形图

描述已自动生成

For other attributes, change the “Binning column” and “aggregation columns” to it’ s attribute before execute it.

Self-reflection.

1. when calculating the number of records and attributes, it would be easier to use the R script to do the counting and easier to calculate the range of attributes.

2. In the bar chart section, it could be better optimized to show the values of all attributes on the same bar chart, with different colours to differentiate them.

The whole workflow for task1 is:

图示, 日程表

描述已自动生成

Problem #1

Task2:

First, we import the required files, “teeth.csv”.

表格

描述已自动生成

In order to report the class labels of the data records at the leaf nodes, I use “RowID” node to rename the name of row, which can makes the name appear on the leaf node.

图形用户界面, 文本, 应用程序

描述已自动生成

Then we use “Hierarchical Clustering”, and put the data into the node, in this case, we define there will be 3 cluster in final.

In the selection of link algorithm, we choose the “Single Linkage”, which defines the distance between two clusters c1 and c2 as the minimal distance between any two points x, y with x in c1 and y in c2.

图形用户界面, 应用程序

描述已自动生成

Then we run this node, we can get a tree like the image below and we can see the name in the leaf nodes clearly.

图表, 直方图

描述已自动生成

If we haven’ t use “RowID” node to rename, we will get a tree with no name in leaf nodes like the picture below.

图表

描述已自动生成

Then we use “Color Manager” to color different cluster(cluster 1, cluster 2, cluster 3).

图形用户界面, 图表

描述已自动生成

Although we have too many attributes and we can only plot in 2D, so I use “PCA” to reduce the dimensionality into 2. And must put PCA dimension to the top 2(PCA dimension 0, PCA dimension 1) by using “Column Restorer”.

图形用户界面, 应用程序

描述已自动生成

图形用户界面

描述已自动生成

Finally, we use “Scatter Plot” node to see how the “Hierarchical Clustering” cluster our data.

图表, 散点图

描述已自动生成

In order to make a simple comparison, I did a control group without categorization.

图示

描述已自动生成

图表, 散点图

描述已自动生成

Now we will see the whole workflow  
