Kierra Young

08/02/2021

DATA 630 -9040

Assignment 5: Classification Unsupervised

Prepared for Professor Firdu

**Introduction**

**Objective**

The objective of this analysis is to use k-means clustering method so that can predict the main factors the determine gram negative bacteria in yeast, The type of analysis that is being conducted in an explanatory analysis and the specific modeling type that will be used is cluster plotting and anomaly detection. With the explanatory analysis and modeling methods of choice the main questions that could be answered are “Does altering the k value influence the outcomes for gram negative bacterial yeast?” and “ What is an ideal distance between sums squared for detecting gram-negative bacteria in yeast?

**Problem Domain**

The background information for problem domain of this dataset is the dataset is based on yeast. The goal of the study was to find out the rules for “predicting protein localization sites in Gram-negative bacteria, given the amino acid sequence information alone”(). According to the article, some interesting statistics are below:

“We considered four localization sites: the cytoplasm, the inner (cytoplasmic) membrane, the periplasm, and the outer membrane. Most rules were derived from experimental observations. For example, the rule to recognize an inner membrane protein is the presence of either a hydrophobic stretch in the predicted mature protein or an uncleavable N-terminal signal sequence. Lipoproteins are first recognized by a consensus pattern and then assumed present at either the inner or outer membrane. These two possibilities are further discriminated by examining an acidic residue in the mature N-terminal portion. Furthermore, we found an empirical rule that periplasmic and outer membrane proteins were successfully discriminated by their different amino acid composition. Overall, our system could predict 83% of the localization sites of proteins in our database.”( M;, N. K. K.)

**Method Rationale**

The method of rationale that will be used for understanding this data will be by using K-means clustering. By using k-means clustering method the method will group all the data instances that are common in the yeast dataset. All the data in the vehicle data set will be grouped by similarities and differences from those characteristics. It is applicable for the problem at hand because it can easily identify the variables that will affect one having certain gram-negative bacterial yeast in their system. It will be able to identify this by calculating the sum of squared distances between clusters increases and the sum of squared distances within clusters decreases for finding the higher k values.

**Analysis**

**Data**

The characteristics of this dataset are listed in the table below. The table below displays all of the characteristics, attributes, associated tasks, instances, missing values, area, and date the dataset was collected and acquired.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 1484 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 8 | **Date Donated** | 1996-09-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 333894 |

(UCI machine Learning Repository)

To continue, some information about this dataset is that the nature of it comes from Creator and Maintainern .Kenta Nakai, of Institue of Molecular and Cellular Biology in Osaka, University Japan. The dataset has been used in the article Predicted Attribute: Localization site of protein.

“The references below describe a predecessor to this dataset and its development. They also give results (not cross-validated) for classification by a rule-based expert system with that version of the dataset. Reference: "Expert Sytem for Predicting Protein Localization Sites in Gram-Negative Bacteria", Kenta Nakai & Minoru Kanehisa, PROTEINS: Structure, Function, and Genetics 11:95-110, 1991. Reference: "A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells", Kenta Nakai & Minoru Kanehisa, Genomics 14:897-911, 1992.”( M;, N. K. K.)

To explain the variables a detailed description is below:

“Attribute Information:

1. Sequence Name: Accession number for the SWISS-PROT database

2. mcg: McGeoch's method for signal sequence recognition.

3. gvh: von Heijne's method for signal sequence recognition.

4. alm: Score of the ALOM membrane spanning region prediction program.

5. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.

6. erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.

7. pox: Peroxisomal targeting signal in the C-terminus.

8. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.

9. nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.”

(UCI machine Learning Repository)

**Exploratory Analysis**: perform exploratory analysis, leverage functions such as “str” and “summary” and discuss their outputs. Also, select few key variables (including the target variable for supervised learning) and study their distributions using plots such as histograms, box plot, bar chart, etc.

Str & Summary

The str describes the data frame and in the data frame there are 1484 observations and 9 variables. The output is below. All the values are numerical except for class variable which is a

character variable. The summary category gives the min, max, 1st/3rd quarters, mean and median.

The output is below under the str output.

str(seeds)

'data.frame': 1484 obs. of 9 variables:

$ mcg : num 0.58 0.43 0.64 0.58 0.42 0.51 0.5 0.48 0.55 0.4 ...

$ gvh : num 0.61 0.67 0.62 0.44 0.44 0.4 0.54 0.45 0.5 0.39 ...

$ alm : num 0.47 0.48 0.49 0.57 0.48 0.56 0.48 0.59 0.66 0.6 ...

$ mit : num 0.13 0.27 0.15 0.13 0.54 0.17 0.65 0.2 0.36 0.15 ...

$ erl : num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...

$ pox : num 0 0 0 0 0 0.5 0 0 0 0 ...

$ vac : num 0.48 0.53 0.53 0.54 0.48 0.49 0.53 0.58 0.49 0.58 ...

$ nuc : num 0.22 0.22 0.22 0.22 0.22 0.22 0.22 0.34 0.22 0.3 ...

$ class: chr "MIT" "MIT" "MIT" "NUC" ...

> summary(seeds)

mcg gvh alm

Min. :0.1100 Min. :0.1300 Min. :0.21

1st Qu.:0.4100 1st Qu.:0.4200 1st Qu.:0.46

Median :0.4900 Median :0.4900 Median :0.51

Mean :0.5001 Mean :0.4999 Mean :0.50

3rd Qu.:0.5800 3rd Qu.:0.5700 3rd Qu.:0.55

Max. :1.0000 Max. :1.0000 Max. :1.00

mit erl pox

Min. :0.0000 Min. :0.5000 Min. :0.0000

1st Qu.:0.1700 1st Qu.:0.5000 1st Qu.:0.0000

Median :0.2200 Median :0.5000 Median :0.0000

Mean :0.2612 Mean :0.5047 Mean :0.0075

3rd Qu.:0.3200 3rd Qu.:0.5000 3rd Qu.:0.0000

Max. :1.0000 Max. :1.0000 Max. :0.8300

vac nuc class

Min. :0.0000 Min. :0.0000 Length:1484

1st Qu.:0.4800 1st Qu.:0.2200 Class :character

Median :0.5100 Median :0.2200 Mode :character

Mean :0.4999 Mean :0.2762

3rd Qu.:0.5300 3rd Qu.:0.3000

Max. :0.7300 Max. :1.0000

Bar Chart

For the bar chart in figure 1 in the appendix, it shows all the classes of yeast in the dataset. The reason I chose these metrics as a bar chart is because the analysis will be examining how many times the instances of independent variables are showing when it comes to the dependent variable ‘class’. From looking at this bar chart, NUC and CTY have a lot of findings whereas the other variables are not as prevalent this means that NUC or CTY might also contain of a lot of gram-negative bacterial instances as we get into data modeling.

Histogram

In the histogram I decided to use the yeast variable MIT and MIT is the “score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins”( UCI Machine Learning Repository). I wanted to view this information because as we get into the cluster analysis and anomaly detection part this variable seems to show even more than CTY. From the diagram which shows the percentage of years and the weight of the grams, the highest weight of gram is around 0.2 and the percentage of the weight is nearing 25%. This is interesting because with MIT being a score of the amino acid content of two different types of proteins it would seem like it would have a reason to show up more in the cluster analysis which will be discussed later.

**Preprocessing**

For pre-processing the yeast data set, there were no missing values, and all the values were numerical, so the histogram and bar chart could be made with no pre-preprocessing. Before removing the class variable, I made the histogram and stacked bar chart to view the yeast data by class and other factors. For pre-processing I also decided to scale the data variables in the dataset. I used a command for the data to be read as ‘seed’ to remove the variable class. The command is below.

myyeast<-seeds

myyeast$class<-NULL

head(myyeast)

**Algorithm Intuition**

The intuition for the K-means method algorithm is to be able to see the instances of the data in cluster groups. Instead of the data having labels they are all put into cluster groups so one can view where most of the data is going. The key inputs of the parameters will be 4 and 2 iterations these will help us better understand which instances are going into the clusters. Once the algorithm is completed one will be able to know the sums squared by cluster within and the sums squared for cluster between.

**Cluster Development & Anomaly Detection**

For the cluster development I decided to use a code that would store the yeast data into the input kc for making the k-means clustering and split the data into 4 and 2 iterations. I chose the parameters 4 and 2 because I wanted to examine the differences between the sums squared betweens and the sums squared within. This will show which model is going to preform better, one with many clusters or one with not that many. For the anomaly detection I used a code that could detect the outliers for both of the 2 iterations that I made. With the outliers I would know the instances that were most unlikely to come up or be the cause of gram-negative bacterial yeast. In the code, I used the parameter 15 for the center in the cluster then used an equation to calculate the distances.

**Result**

**Output**

From the k-means clusters method and the anomaly detection the output of the clusters is in the appendix as figure 3 and figure 4. First model shows the cluster with 4 iterations and in this model, it displays all the instances that are common to each separate cluster. According to the cluster it says that the points 990 is in cluster 4 and 503 is in 3 and so on for the other clusters. Likewise, the second cluster with two iterations only show two separate clusters. There are also many different instances in that box as well all the clusters.

**Model Propertie**s

Differences between iteration 4 and 2 and is how they were created. For example, for the clusters with 4 iteration it grouped the iterations sizes 459, 602, 182, and 241. There differences are the number of iterations and how they show in the model. For the second cluster plot 2 clusters of sizes 460 and 1024 were grouped into clusters. The first cluster has less groupings while the other cluster has 2 groups of all the instances. One similarity is that the number of instances stay the same and the percentage of 38.56-point variability also stays the same.

**Evaluation**

For the metrics there were two sum squared equations one for iteration 4 and iteration 2. For iteration for within cluster sum squared by cluster equation was 9.242984 23.071549 26.505486 13.363215 (4 iterations clusters) and (between\_SS / total\_SS = 41.5 %). 41.5 percent is the within cluster sum of squares for the 4 iterations. The distance total sum of squared distances is 72.183234 +9.242984= 81.426218. The within distance is 72.183234 and the between distance is 9.242984. For the second iteration the cluster equation was 58.55909 37.27156 (2 iteration clusters) and (between\_SS / total\_SS = 22.4 %). The distance total sum of squared distances is 95.83065 + 58.55909= 154.38974. The within distance is 95.83065 and the between distance is 58.55909. Also for clustering class the main classes with the highest instances were MIT 169 75, NUC 51 378, CYT 78 385, and ME3 24 139. For MIT had the highest amount of instances in 1 as 169 while had NUC as the highest amount in 2 as 378.

**Conclusion**

**Summary**

The key findings of this analysis are that the k-means cluster method was able to detect that MIT and NUC have the highest number of instances amongst the yeast data. It was also able answer the “Does altering the k value influence the outcomes for gram negative bacterial yeast?” and “ What is an ideal k value is for detecting gram-negative bacteria in yeast”, questions in the evaluation section by showing the distance between sums squared, the within sums squared, and the between sums squared values. It was proven that certain k-values can influence detection for gram- negative bacterial yeast especially given the number of clusters in the k-value. In this analysis we only examined two different k-values. One other interesting finding is that NUC was one of the highest classes in the dataset amongst class and is also high in the class cluster analysis.

**Limitations**

Some limitations in this analysis were that there was a lot of data in this dataset so viewing the different clusters could be a little hard to interpret. So, I decided to use different parameters so that viewing the instances would be easier to read or see. Also, the k-value equation and figuring out how to calculate the distance between sums was a lot difficult for interpretation since I was not sure if those just integers of instances or actual percentages.

**Improvement Areas**

Some areas that I can improve in are when it comes to interpreting the quantitative measures of the data such as within, between, and distances of the sums squared. Likewise, when it comes to setting parameters, I could have tried a larger K-means cluster number instead of just 2 and 4, so that I could see what would happen on a larger scale.

Appendix

Figure 1. Bar Chart

Chart, bar chart

Description automatically generated

Figure 2. Histogram

Chart

Description automatically generated

Cluster 4

Diagram, bubble chart

Description automatically generated

Cluster 2

Diagram

Description automatically generated with medium confidence

References

M;, N. K. K. (n.d.). *Expert system for predicting protein localization sites in gram-negative bacteria*. Proteins. https://pubmed.ncbi.nlm.nih.gov/1946347/.

UCI machine Learning Repository: Yeast data set. (n.d.). http://archive.ics.uci.edu/ml/datasets/Yeast.