IST707 - Data Analytics

Final Project Report

Nianyou Zhang

12/4/2020

**Heart Failure Prediction**

The purpose of this project is to analyze a set of heart failure data using several data analysis algorithms, from data exploration to prediction. The report has five sections with distinct objectives: introducing dataset, exploring the dataset, recognizing the limitation of the dataset, building predictive models, and finally concluding recommendations.

**Dataset introduction:**

The heart failure clinical record is found on Kaggle.com (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data). The dataset contains 13 attributes, 299 observations. The real-world problem is to firstly, understand what pre-conditions and blood-test results are related to heart failure; secondly, build predictive models, with input pre-conditions and blood-test results, to help patients predict potential heart emergencies.

The attributes of this dataset can be categorized as following:

|  |  |  |
| --- | --- | --- |
| Basic information: | | |
| Attribute name | age | sex |
| Data type | num | num, binary |
| Description | age of the patient | 0 for female, 1 for male |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pre-existing conditions: | | | | |
| Attribute name | smoking | diabetes | high\_blood\_pressure | anaemia |
| Data type | Boolean | Boolean | Boolean | Boolean |
| Description | if the patient smokes or not | if the patient has diabetes | if the patient has hypertension | decrease of red blood cells or hemoglobin |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Blood-test results: | | | | | |
| Attribute name | platelets | diabetes | high\_blood\_pressure | anaemia | ejection\_fraction |
| Data type | num | num | num | num | num, percentage |
| Description | platelets in the blood (kiloplatelets/mL) | level of the CPK enzyme in the blood (mcg/L) | level of serum creatinine in the blood (mg/dL) | level of serum sodium in the blood (mEq/L) | percentage of blood leaving the heart at each contraction |

The target variable in the dataset is DEATH\_EVENT (boolean). It records if the patient deceased during the follow-up period. 96 out of 299 observations have deceased from heart failure in this dataset.

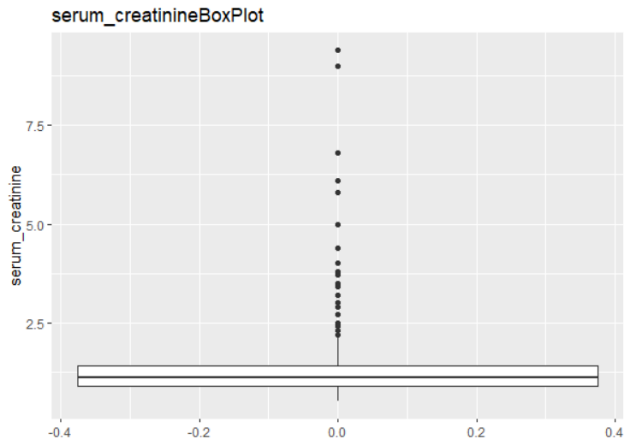
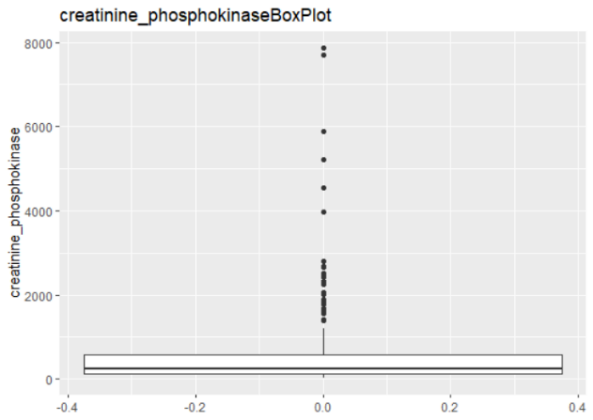
The attribute “time (num, days)”, that records the follow up period, is not used in my analysis.

For this dataset, the data preprocessing work mostly depends on the specific task, because this data is generally good: no missing value and columns are in correct data types. The only general preprocessing step I did was to drop “time” column, and the reason is that the “follow-up period” does not affect patients’ health condition.

For association rule mining, I binned numeric data according to their distributions (using quartiles) and I turned Boolean and binary columns into factor columns. For clustering, SVM, ANN, I used min-max method to scale numeric data.

**Dataset exploration:**

I firstly created box plots for all the numerical data to visualize their distributions. Creatinine columns tend to have extreme outliers (as shown below), however, most other numerical columns are normal. Categorizing these numerical columns using quartiles will include all the outliers in one category (if outlier condition is dangerous, they will at least be in the same “dangerous” category), thus will be fine for analysis.



After turning all columns into factors, I created transactions for all observations, getting ready for association rule mining. By looking at item frequency table, I noticed the sex for dataset is disproportioned: 35% female, 65% male. This information is important when considering the bias in association rules, because “male” has larger number than female in this dataset, but naturally male and female population should be close to 1:1.

I saw that many of the precondition columns are similarly disproportioned. Since the dataset only includes patients’ data (no normal people), it is not surprised to see disproportioned preconditions. However this disproportion can create bias for association rules.

After a trying out a range of support and confidence values, the following combination yield most appropriate number of rules.

* Support=0.05;
* Confidence=0.7;
* Minlen=4;

This combination results in 10 rules. The best rule is:

|  |  |  |
| --- | --- | --- |
| Rule: | LHS | RHS |
| ejection\_fraction=1st | DEATH\_EVENT=Yes |
| serum\_creatinine=4th |
| serum\_sodium=1st |
| Parameters: | Support | 0.05 |
| Confidence | 0.79 |
| Lift | 2.46 |

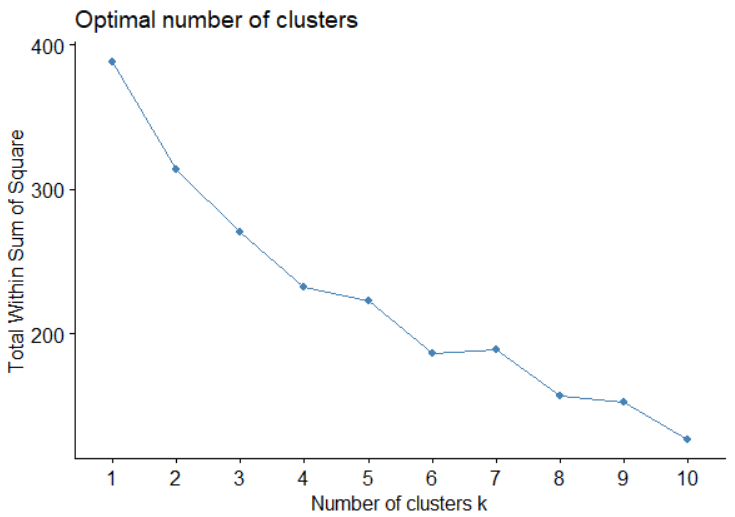
The confidence of 0.79 is quite strong, because 79% patients with such attributes passed away. The lift of 2.46 is not extremely high, but high enough to indicate the dependence between the three conditions and death.

**Dataset limitations and the attempt to fix:**

This dataset has a crucial limitation: the target value heart failure death event cannot reflect the patients that are at critical condition and are about to pass away. In another word, patients that are in death-level terrible condition are labeled as alive. This can be very misleading. One solution to this problem is to apply clustering on the dataset without “death event” label. This way, the patients will be categorized only base on their attributes. After clustering, I can then compare the death rate among all the clusters, to discover the significant differences.

To prepare for clustering, all numerical data are scaled; Boolean and binary factors are made sure to be 0s and 1s.

The first step for k-means clustering is to find k value with elbow plot:



As the graph shows, the decrease of sum square error in clusters starts to flatten when K equals 4, thus I should create four clusters.

The four clusters and their respective death rate are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster #: | 1 | 2 | 3 | 4 |
| Death rate: | 0.384 | 0.296 | 0.290 | 0.338 |

The total sum of squares: 388.5571, given 299 observations. This means the 4 clusters are fairly accurate and significant.

By looking at the summary statistics for cluster 1 and 3, I noticed that cluster 1 patients all have anaemia while cluster 3 patients all do not have anaemia. Cluster 3 patients have almost twice as high creatinine phosphokinase level than cluster 1 patients.

This result can indicate that anaemia and low creatinine phosphokinase level are significant risk factors for heart failure death. However, this result should be looked at critically because the clusters are created to be as different from each other as possible.

**Predictive models:**

The last part of this project is to build a predictive model to evaluate heart failure risk given patients’ pre-conditions and blood test results.

Given that the dataset is consist of all numerical data, and the previous clustering result is fairly successful, SVM is one of the best fitting choice for building predictive model. At the same time, previous data exploration showed the complexity of the attributes and target label, the best predictive model should be ANN because ANN’s backpropagation will help algorithm adapt to the data when data has no clear pattern. In the real world, ANN is the best choice because it also adapts to continuous learning. For comparation, I will also run classic decision tree model alongside ANN and SVM. All models will be performed with repeated cross validation (5 sets, repeated 5 times), since none of the three models have built-in cross validation like ensemble methods do.

For SVM, I performed linear boundary, polynomial boundary and radial. All three algorithms are performed with a range of parameters to obtain the best result. The results are as follow:

|  |  |  |  |
| --- | --- | --- | --- |
| Type: | SVM linear | SVM polynomial | SVM radial |
| Test Accuracy | 0.8305 | 0.7288 | 0.7458 |
| Train Accuracy | 0.7625 | 0.7792 | 0.8292 |
| Specificity | 0.8462 | 0.7143 | 0.6667 |
| Precision | 0.95 | 0.95 | 0.9 |
| Recall | 0.7755 | 0.8837 | 0.8182 |
| F\_measure | 0.8539 | 0.9157 | 0.8571 |

Test accuracy and recall are most relevant to heart failure diagnose. Train accuracy is viewed critically because a high training accuracy can indicate overfitting. Here the preferred SVM model is polynomial.

For ANN, since the dataset attributes can be categorized into three categories, the hidden layers should have around 3 nodes. In addition, because the dataset is relatively small, a slightly large decay is preferred.

Last but not the least, decision tree is added to compare with SVM and ANN. For decision tree, the binary and Boolean attributes in data are kept as categorical, while numerical attributes are left as numerical without scaling. The algorithm will determine where to cut the numerical columns.

Here are the detailed evaluation measures for all three models:

|  |  |  |  |
| --- | --- | --- | --- |
| Type: | ANN | SVM polynomial | DT |
| Test Accuracy | 0.8136 | 0.7288 | 0.6441 |
| Train Accuracy | 0.7542 | 0.7792 | 0.7792 |
| Specificity | 0.7857 | 0.7143 | 0.3333 |
| Precision | 0.925 | 0.95 | 0.9 |
| Recall | 0.7708 | 0.8837 | 0.9474 |
| F\_measure | 0.8409 | 0.9157 | 0.9231 |

For this dataset, model accuracy and recall are both important to look at. Recall is important because in diagnosing patients, it is safer to have false positive than false negative. From above evaluation measures, ANN has best testing accuracy, and good overall measures. When I take in consideration of training set accuracy, ANN is much less biased than SVM. The conclusion for predictive model is: Artificial Neural Network is a best choice.

**Recommendation:**

Blood test results are strong indicators of fatal heart conditions. From clustering analysis, patients with decrease of red blood cells or hemoglobin have nearly 9% increased heart failure risk. From association rule mining, blood with high serum creatinine and low serum sodium are closely associated with heart failure. For people that have concern over their heart condition, they should take blood test regularly to monitor the key parameters mentioned above.

For clinic, when new patients register in the clinic system, their pre-existing conditions and blood test result should be entered into ANN predictive model, to evaluate their heart failure risks.

Work Cited

datasets\_727551\_1263738\_heart\_failure\_clinical\_records\_dataset.csv Heart Failure Prediction. https://www.kaggle.com/andrewmvd/heart-failure-clinical-data