CS777 – Term Project

Gaurav Tungare

*Oct 18, 2021*

1.Objective

|  |  |
| --- | --- |
| **Classification** | **Clustering** |
| An opportunity to build and test a classifier model using a real-world data. The data to be used is a part of the **2018 BRFSS Survey Data prepared by CDC**.  I will trying to predict “that the person was ever told to have some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia” | Learning algorithm for **Gaussian Mixtures using the expectation-maximization algorithm**. I have used gap minder dataset To gather insight into major trends and patterns for World Health and Economics, also identify the specific global statistical trends that have not reached a broad public audience |

**Classification**

2. Background of data - 2018 BRFSS Survey Data prepared by CDC



The project-2018-BRFSS-arthritis.csv file has the dataset for the project and it has 11933 tuples and 108 attributes. Each tuple is a person who participated in the survey and each. attribute is an answer to a survey question. The class attribute is havarth3 and its value is

either 1 or 2. The value of 1 means that the person was ever told to have some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia.

* **Total Record Count – 11933 tuples w/ 108 attributes**

After Train and Test split between dataset

* Total **“Train”** Record Count **- 7997**
* Total **“Test”** Record Count **- 3936**

3. Names of all classifier algorithms used in the **project**

For each attribute selection method below classification algorithm were run

* Logitic Regression
* Decision Tree
* RandomForest
* Naive Bayes

5. Test results of all 4 models

Text

Description automatically generated

6. Dimensionality Reduction

There were 107 attribute which were selected in the original model. As stated in the class, we do apply dimensionality reduction to see what attribute we have to select. Now generally we apply various technique for reduction/selection

I applied dimensionality selection technique to reduce about 30% of dimension

7. Test results of all 4 models- After Dimensionality reduction

Text

Description automatically generated

8. Best Model

The best model after 100 iteration is **RandomForest** classification gave the best performance across

**=== Detailed ===**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy| | F1Score| | Precision | Recall |
|  | 0.75187 | 0.73239 | 0.747328 | 0.7518700 |

Justification

The above model has **Accuracy** of 0.7518 which indicate that the model is able to correctly predict the value of some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia, TP rate ~0.80 indicate a good performing model

The **F-measure** of 0.732 indicate good sensitivity and recall for over all model, the close the F-measure to 1 the prediction across test data set improves, again this measure gives us confidence that the class attribute which the model is predicting is good

9. Five attributes most relevant to the class attribute

The five attribute which are most relevant to class attribute which helped predict the correct value for some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia

1. Variable Name: **CHCCOPD1** - you have chronic obstructive pulmonary disease, C.O.P.D., emphysema or chronic bronchitis. this indicates a chronic health condition
2. Variable Name: **DIFFWALK** - Difficulty Walking or Climbing Stairs

This column indicates overall health of the person predicting the overall well being. If the person has answered yes to the question it would mean underlying health condition

1. Variable Name: \_**RFHLTH** Adults with good or better health

This column indicates overall health of adults with good or better health, which will drive the predicting if the person has some form of issues with health or not

1. Variable Name: \_ **AGE65YR** Reported age in two age groups calculated variable

This column has three categories of age 18 to 64 or Age 65 or older , this is important column as it will drive other columns to predict the health attribute correctly

1. Variable Name: \_ **EMPLOY1** Employment Status

Surprisingly, this column which indicates relation between if person is employed and presence/absence of disease. Your employment status does drive overall health of the person which in turn helps predict the if certain disease if present/absent

10. Lessons learned from this dataset .

There are various aspect of data mining which I have learnt from this project.

Let me summarize on each of the stuff

From Data cleaning

Before performing any data analysis, the available data has to be cleaned to remove as many errors as possible. The goal is to convert the available data into high quality but in our case data cleaning has not much effect on the classification

Classification

Learnt various aspect of classification on various classification algorithms

Naïve Bayes, Logistic, Decision Tree & Randomforest

On the test dataset the model performance is not varying a lot , the deviation across each model is +/- 3% which indicate that you need to be very careful with attribute selection method

Model Performance

I learnt about general prediction ability of a model all 5 models were analyzed ,various accuracy, sensitivity, specificity, precision, F measure (or F-score). These different measures represent different aspects of the performance of a model which was used to predict some form of

arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia.

**Clustering**

11. Background of data – gapminder data set

A selection of world health and economics statistics from the Gapminder project can be found in the dslabs package as data(gapminder) and python pip install gapminder

“[*The Best Stats You've Ever Seen*](https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen?language=en) “ :-The original Gapminder TED talk

<https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen?language=en>

<https://www.gapminder.org/>

* **Total Record Count – 1704 tuples**

Year Range- 1960 to 2016

* Country.
* Year.
* Life\_expectancy. Life expectancy in years.
* Population. Country population.
* GDP. GDP according to World Bankdev.
* Continent.

Text

Description automatically generated

After Train and Test split between dataset

* Total **“Train”** Record Count **- 1123**
* Total **“Test”** Record Count **– 581**

12. Analysis of data – gapminder data set

From above scatter plot of population gdp and the Life expectancy , we can definitely see clustering around age70 years also around age 40 yrs well

Chart, scatter chart

Description automatically generated

13. Clustering using Kmean

I am using Silhouette Score for measuring the goodness of the clusters

Please note this is used a measure to see the clustering

Text

Description automatically generated

Chart, line chart

Description automatically generated

14. Clustering using Gaussian Mixtures using the expectation-maximization algorithm.

I am using Gaussian Mixture to demonstrate the clustering technique learned in the class, The gaussian mixture using EM algo does give the required clusters

A picture containing text, plaque

Description automatically generated

Lessons learned from this dataset .

In general, k-means and EM may perform better or worse, depending on the nature of the data we want to cluster, and our criteria for what defines a good clustering result.

Thanks