INTRODUCTION TO AI: DATA ANALYSIS USING SUPERVISED AND UNSUPERVISED METHODS

Shalev Habany
August 2021

ABSTRACT

In this paper analyzed 3 different algorithms of supervised learning and 3 different algorithms of unsupervised learning.
The algorithms were applied on the mushrooms data set and deleted mushrooms data set, displayed results, and analyzed the performance in attempt to find the most suiting methods for each learning approach.
I found that the supervised methods have the best performance of classify the data set, it is also the reason I chose the supervised methods to classify the deleted data instead of the clustering.
There is also a compare between performance before and after dimension reduction.
We will see that the clustering was significantly better after dimension reduction algorithm was applied however, the supervised classification methods were not affected and even got worse when dimension reduction applied on the data set. All source code can be found in this GitHub repository: `https://github.com/shalev-habany/Introduction-to-AI`

1 INTRODUCTION

First, I will introduce the task and the approaches in order to solve the task.
There are two data sets, ”mushrooms data set” and ”deleted mushrooms data set”, the mission is to classify the data sets when the target feature (labels) is the odor feature.
Displayed here 2 approaches on the mushrooms data set and 1 approach on the deleted mushrooms data set in order to classify them.
The approaches which we used for the ”mushrooms data set”:
First approach: Apply a clustering algorithm on the data set and compare the clustering to the real labels (odor feature).
Second approach: Apply a supervised classify algorithm on the data set and comparison the quality of the methods by using measures for supervised learning algorithms.
There is also compare between the quality of each approach before and after dimension reduction (which described at the mission as ”third approach”).
For the ”deleted mushrooms data set” I used only the second approach .

1.1 THE MUSHROOMS DATA SET

The main data set of this article is the ”mushrooms data set”, a data set which contains 6499 mushroom samples.

1.1.1 ATTRIBUTES INFORMATION

Each sample described be 22 categorical attributes (features), 18 of them are multi-class (more than 2 categories) attributes and 4 attributes which are binary-class (2 categories) attribute. The attributes are summarised below:

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s.

2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color:brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. gill-color:black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10. stalk-shape: enlarging=e,tapering=t
11. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
12. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-color-above-ring:brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
14. stalk-color-below-ring:brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15. veil-type: partial=p,universal=u
16. veil-color: brown=n,orange=o,white=w,yellow=y
17. ring-number: none=n,one=o,two=t
18. ring-type:cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
19. spore-print-color:black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
20. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
21. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d
22. Class: edible=e, poisonous=p

1.2 Handle Categorical Data

Every machine learning algorithm can work only with a numerical data, as we noticed at the 1.1.1 section all of our attributes in the mushrooms data and in the deleted mushrooms data are strings. In order to apply machine learning methods on the data, it needs to be converted to numerical data. So that, I used method which called ”one hot encoding” to convert the data. The one hot encoder convert every feature to sub features (the number of the sub features is the number of the categories of the feature) and matching the feature a binary array. For example: if there is a sample with odor a the one hot encoded feature we look like this:

One Hot Encoded odor

	$odor_a$	$odor_l$	$odor_c$	$odor_y$	$odor_f$	$odor_m$	$odor_n$	$odor_p$	$odor_s$
sample odor	1	0	0	0	0	0	0	0	0

We will treat every sub feature like this as feature which means the mushrooms data set contains 110 features (every category is feature).

1.3 The Deleted Mushrooms Data Set

The deleted mushrooms data set has same attributes as the regular mushrooms data set have however, there are 1625 mushroom samples and part of the samples have deleted attributes (represented as ”-”).
I handled that with the one hot encoding (described at 1.3 section).
if there was ”-” I made every sub feature of the feature equals 0.

2 Methods

All of the methods in this section were used from these python packages:

- scikit-learn
- tensorflow

2.1 Dimension Reduction

I used here PCA dimension reduction twice, once for learning and another time to visualise. Notice that I used this that only with the regular mushrooms data set (not the deleted one)
I used PCA twice:

- For data visualisation I used PCA with n_components=2, keeping 42% of the data variance.
- For learning data I used PCA with n_components=28, keeping 92% of the data variance.

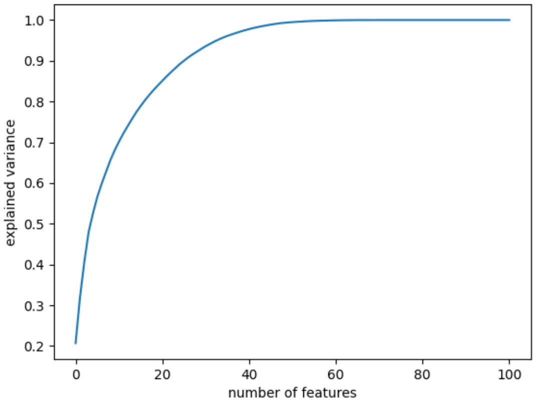

Figure 1: PCA:number of features vs. explained variance

I chose 28 components to learning because as we see it exactly before the explained variance decreased a lot.

2.2 Clustering (Unsupervised Learning)

Notice: I set n_clusters=9 because there are 9 categories of odor.

- K-Means:
  - On the mushrooms data set, **n_clusters**=9.
  - On the mushrooms data set with dimension reduction, **n_clusters**=9
- Hierarchical Agglomerative Clustering:
  - On the mushrooms data set, **n_clusters**=9.
  - On the mushrooms data set with dimension reduction, **n_clusters**=9
- General Mixture Model (GMM):
  - On the mushrooms data set, **n_clusters**=9.
  - On the mushrooms data set with dimension reduction, **n_clusters**=9

2.2.1 Clustering Quality Measures

For measure the data separation quality I used the **mean_silhouette_score**.
For measure the similarity between the clusters and the ground truth, I used the **fowlkes_mallows_score**.

2.3 CLASSIFICATION (SUPERVISED LEARNING)

Notice: I used a different approach for the supervised learning.
Instead of build one model that classify the target feature (odor), I related to every sub feature of the odor feature (sub feature defined at 1.3 section) as regular feature and built 9 models that predict a binary class instead of one model that predict multi class

- Linear SVM (Used SVC with linear kernel method from scikit-learn):
  - On the mushrooms data set, default parameters and linear kernel.
  - On the Deleted mushrooms data set, default parameters and linear kernel.
- Random Forest:
  - On each data I tried from 1 to 20 estimators and took the one that gave the best mean accuracy score.
  - On the mushrooms data set, **n_estimators=10**.
  - On the mushrooms data set with dimension reduction, **n_estimators=10**.
  - On the Deleted mushrooms data set, **n_estimators=20**.
- Neural Networks:
  - The optimizer parameters: **optimizer=”adam”** (adaptive optimizer), **loss function=”binary cross-entropy”**, 50 epochs
  - On the mushrooms data set:
    - ∗ Input layer: number of neurons is the number of the sub features (regular data set 101, after dimension reduction 28).
    - ∗ 2 hidden layers: 128 neurons and ”relu” activation function.
    - ∗ Output layer: one neuron and sigmoid function (binary classification).
  - On the Deleted mushrooms data set:
    - ∗ Input layer: number of neurons is the number of the sub features (13 sub features).
    - ∗ hidden layer: 32 neurons and ”relu” activation function.
    - ∗ hidden layer: 64 neurons and ”relu” activation function.
    - ∗ Output layer: one neuron and ”sigmoid” function (binary classification).

2.3.1 CLASSIFICATION QUALITY MEASURES

For measure the classification quality I used **mean accuracy,** i.e. for each sub feature I measured the accuracy and then took the average.

3 RESULTS

The result which presented in this section will be addressed for 3 data sets.
The regular mushrooms data set with and without dimension reduction have results of the supervised and the unsupervised methods (described at section 2).
In addition, we have also the deleted mushrooms data set that has results of supervised learning methods only (because the supervised learning worked much better on the regular data set, as will be described below, I chose these methods instead of the clustering in order to classify the data optimally)

3.1 CLUSTERING

After preprocessed the data, I applied on the data 3 clustering methods: K-Means, GMM and Hierarchical Agglomerative Clustering. For each method I used 9 clusters because we have 9 categories of odor.
marginal notation:

- I didn’t take the optimal number of clusters to separate the data.
  I preferred to take more than 9 clusters, show the distribution of the labels in every cluster by histograms and then determine which clusters are contained in each odor category. However, it could do the article longer and complicated so I stayed with the 9 cluster approach.

3.1.1 Quality Measure Results

The unsupervised quality measures (2.2.1 section) of the clustering methods:

Clustering Quality Measure

clustering method	mean_silhouette_score	fowlkes_mallows_score	with/without PCA
K-means	0.229	0.577	without PCA
K-means	0.54	0.311	with PCA
GMM	0.161	0.53	without PCA
GMM	0.034	0.635	with PCA
Hierarchical Agglomerative Clustering	0.273	0.638	without PCA
Hierarchical Agglomerative Clustering	0.51	0.33	with PCA

3.1.2 Visualisation

As described at 1.2 section I used PCA with **n_components=2** in order to visualize the data. Notice that the PCA kept only 42% of the data variance so these figures won't be the best measure however, we get some information about how the data looks.

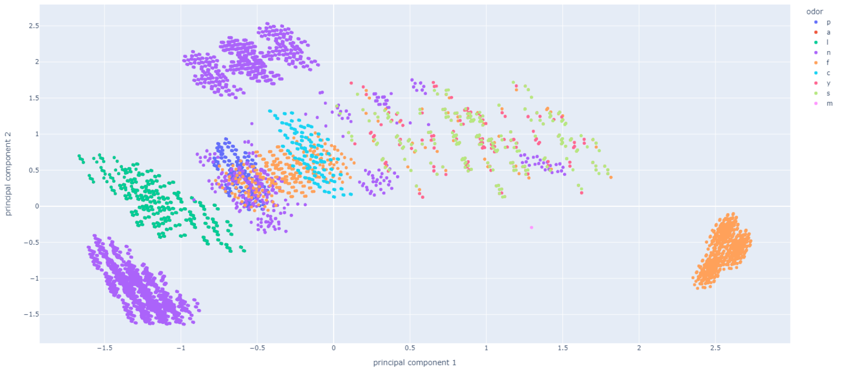

Figure 2: Labels Figure (draw with PCA set n_components=2)

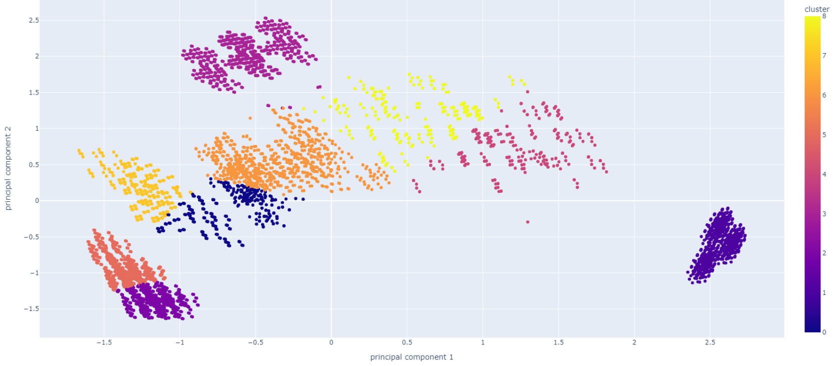

Figure 3: K-Means Clustering Figure (draw with PCA set n_components=2)

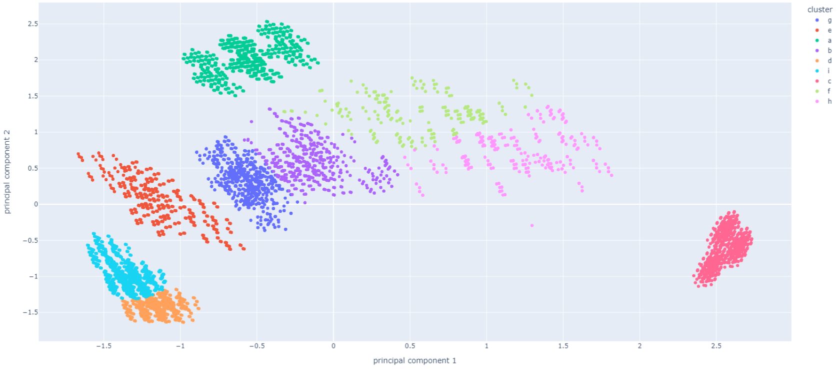

Figure 4: GMM Clustering Figure (draw with PCA set n_components=2)

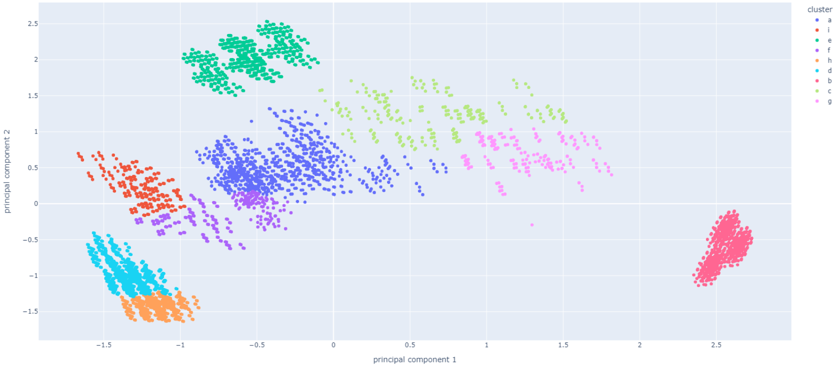

Figure 5: Hierarchical Agglogarative Clustering Figure (draw with PCA set n_components=2)

3.2 Classification

Here the results of the classification methods (described at 2.3 section) on the regular mushrooms data and the deleted mushrooms data are displayed.

3.2.1 Quality Measure Results on Regular Data

The supervised quality measures (2.3.1 section) of classification methods on the regular data:

Classification Quality Measure

classification method	mean_accuracy	with/without PCA
Linear SVM	97.2%	without PCA
Linear SVM	92.8%	with PCA
Random Forest	95.9%	without PCA
Random Forest	96%	with PCA
Neural Network	96.4%	without PCA
Neural Network	97.2%	with PCA

3.2.2 Quality Measure Results on Deleted Data

The supervised quality measures (2.3.1 section) of classification methods on the deleted data (notice that I didn't apply PCA on the missing data) :

Classification Quality Measure

classification method	mean_accuracy
Linear SVM	87%
Random Forest	86%
Neural Network	87%

4 Discussion

4.1 Clustering Approach

Learning on the mushrooms data set, I found that in 3 clustering methods before dimension reduction, there was moderate fowlkes mallows score which means that the similarity to the labels was moderate. After the dimension reduction there was interesting results.

For K-Means and Hierarchical clustering the silhouette score increased which means the separation was better however, the fowlkes mallows score decreased which means the similarity to the ground truth was worst.

For the GMM method it was inverted, the silhoutte score decreased even more which means the data wasn't separated good at all however, the fowlkes mallows increased be 0.1 which means the similarity to the gound truth was better.

In conclusion: The method that had the best results is the Hierarchical Agglogarative clustering.

4.2 Classification Approach

Learning on the mushrooms data, I found that all of the classification methods which described in 2.3 section, classified the data very well.

All of them had above 95% accuracy (with the regular data), the SVM method had the best accuracy also at the deleted data, so we can assume that the data is allowed to linearly separation.

However, the dimension reduction decreased the accuracy for SVM method and slightly increased the other methods.

Surely, the deleted data results was worst than the regular data probably because, there were less samples and less data (deleted data ”-”).

In conclusion: The neural network had almost the same results as the SVM with the both data sets and had the best results with the dimension reduced data set, so it is the best method to classify this data.

4.3 Approaches Comparison

As described in 3 section I found that the supervised methods had much better results to our task and advise the supervised approach more than the unsupervised approach with this data set.