Intelligent Data Analysis Project Report

Han Mingji

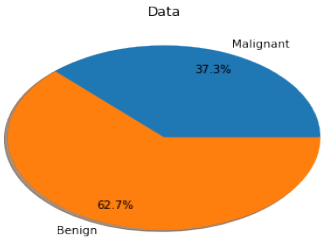
July 21, 2018

1 Abstract

In this project, we use several methods including Principle Component Analysis , Clustering Algorithm and SOM to analyze UCI Machine Learning Breast Cancer Wisconsin (Diagnostic) Data Set.(You can get dataset from `https://www.kaggle.com/uciml/breast-cancer-wisconsin-data`).We can use this dataset to analyze the relationship between components in data. And we use clustering algorithm data on dataset and visualize them. The analysis result can be used to predict the cancer in the future.

2 Data

2.1 Overview



This dataset includes 570 samples. There are 212 mailgnant samples and 357 benign samples.Each sample is 32 features data.

2.2 Features

There are ten features of each samples.Features are calculated from image of a fine needle aspirate (FNA) of a breast mass. They describe features of the cell nuclei in the image. (Feature information are taken from `https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/home`.Thus they are very important components in our dataset.

Number	column	Description
1	radius	mean of distances from center to points on the perimeter
2	texture	standard deviation of gray-scale values
3	perimeter	
4	area	
5	smoothness	local variation in radius lengths
6	compactness	$perimeter^2/area - 1.0$
7	concavity	severity of concave portions of the contour
8	concave points	number of concave portions of the contour
9	symmetry	
10	fractal dimension	"coastline approximation" - 1

The other 20 features are the mean,stand error and largest data of features which are calculated from each feature.The rest of the features: one is the ID .Another is the category of each data (M = malignant, B = benign). All features (except category and ID) are float numbers.There is no missing values in the dataset.

2.3 Represent Data

In order to analyze data, we need to preprocess our dataset so that it can be used for data analysis. We use 30 features of data to create our dataset. The category (malignant, benign) is used for verifying the correctness of our analysis and prediction. The ID is useless in our analysis so we remove it. We create matrix to represent our dataset.

$$X \in R^{d \times N}$$

D is the dimension of our data (d = 30),N is the size of our dataset (N = 570).Each samples are represented as columen vectors.

2.4 Normalization

Since each dimension represents different kinds of data, we need to normalize our dataset.And after normalization, the covariance matrix can show the correlation betweeen each features.If we do not do this, it will be hard for us to analyze our dataset.

For each dimension, we calculate expectation and standard variance.

$$E[x] = \frac{1}{N}\sum_{i=1}^{N} x_i^d$$

$$Var[x] = \frac{1}{N}\sum_{i=1}^{N}(x_i^d - E[x])^2$$

Then we normalize our data by using this formula.

$$x_{new} = \frac{x - E[x]}{Var[x]}$$

3 Analysis

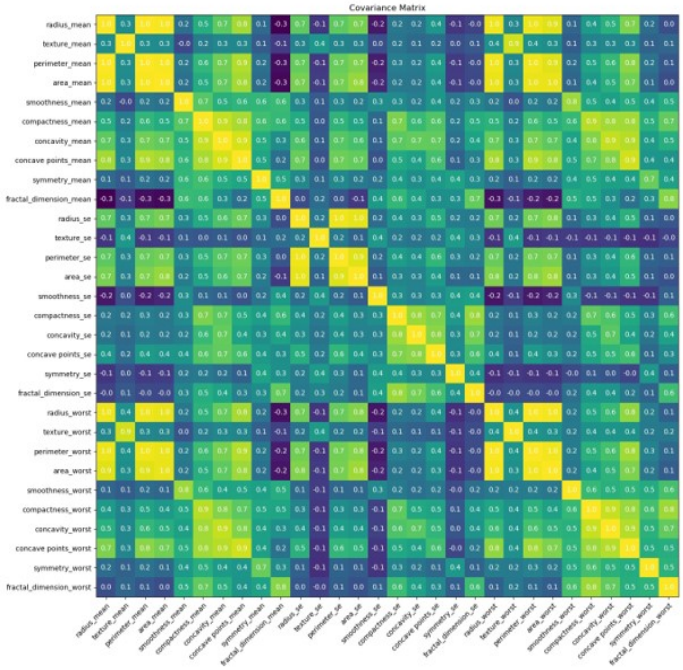
So far, we have an overview and preprocessing on our dataset.Here are some problems that we are interested in:

1. What is the realationship between each features?

2. Can this algorithm find the nature structure of data?Can the clustering result of data represent the real distribution of data ? What is the common point of data in the same category ?

3.1 Covariance Matrix

To find out the realationship between each features we use the data that is normalized to calculate Covariance Matrix.And here is the visualization of the covariance matrix.



$$C = \frac{1}{N} X X^T$$

In this matrix, we find that most features are closely related as most covariance is greater than zero.Only a few covariance is smaller than zero. The close realationship between each dimension shows all these features are important

to the category of cell.And it is a useful feature for dimension reduction. We may preserve most information and only use three or four dimension. It is also reasonale if we consider the situation in real life. The data descibe the feature of cell nuclei. Thus these dimensions are likely to be related closely.

3.2 Principle Component Analysis

So far we get the dataset which is normalized and covariance matrix.To solve the second question and validate whether our hypothesis is correct or not. We use principle component analysis for dimension reduction.

First we get the eigenvectors matrix of covariance matrix C.Each eigenvector is a column matrix.

$$Cv = \lambda v$$

We can get new covariance matrix. The new covariance matrix is a diagonal maxtrix. As PCA eliminates the interplay between each dimension and and make data in new space more aligned.

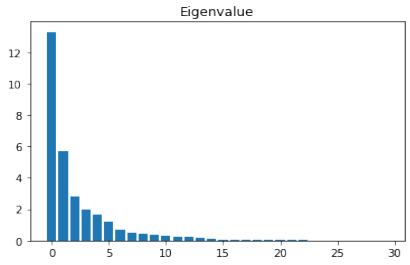
$$C^{'} = V^{T}CV$$

The new dataset can also be calculated.

$$X^{'} = V^{T}X$$

But in the new space, we should not preserve all dimensions. Now we analyze our eigenvectors and eigenvalues of matrix to decide how many dimensions and which dimension should we preserve.

3.3 Eigenvalues And Eigenvectors

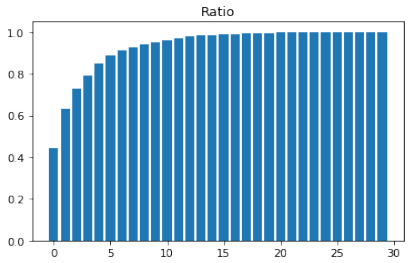


3.4 Eigenvalues And Eigenvectors

We sort the eigenvalues in descending order.It is obvious that the first two eigenvalues are much larger than the others. This means that the key information is preserved in these two dimension. We calcuate amount of variance

explained by using this formula to determine how many axes should we use for clustering and visualization:

$$\rho_t = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{d} \lambda_j} (i = 1, 2, 3...., d)$$

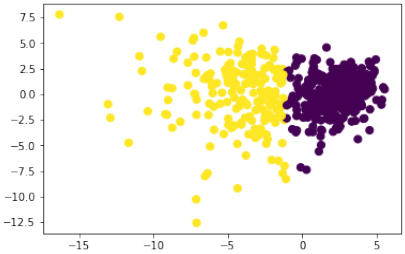


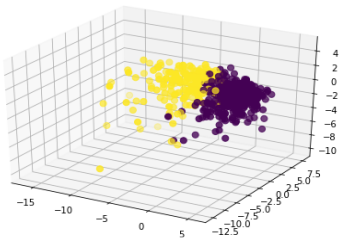
If we project our original data onto 3 dimension space whose axes' directions are the eigenvectors with the three greatest eigenvalues. We preserve 76% variability of data.This means we can reduce the dimension of data to two or three dimensions without lossing too much information.

3.5 Clustering

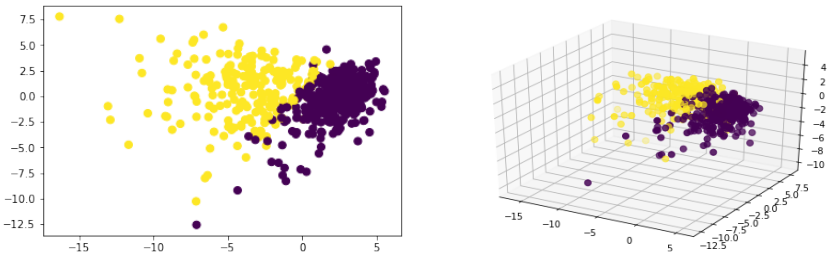
The clustering algorthim picks data centers and get clustering results after many iterations. It is an unsupervised learning algorithm without any priror knowlegde.We hope to use clustering algorithm to find the realationship between different type of data. If the clustering algoritm can sepearate our dataset successfully, we indicates that the clustering algorithm can find the nature relationship between different types of data by using information we have preserved. For our cancer dataset, good clustering can help us to find the boundary between two category and this will be helpful for prediction when we have new data.

We use online clustering algorithm and set clusters as two to see whether the clustering result can find the nature of data. We use this algorithm on two and three dimensions data which we preserved by using PCA. Following images show the clustering result.





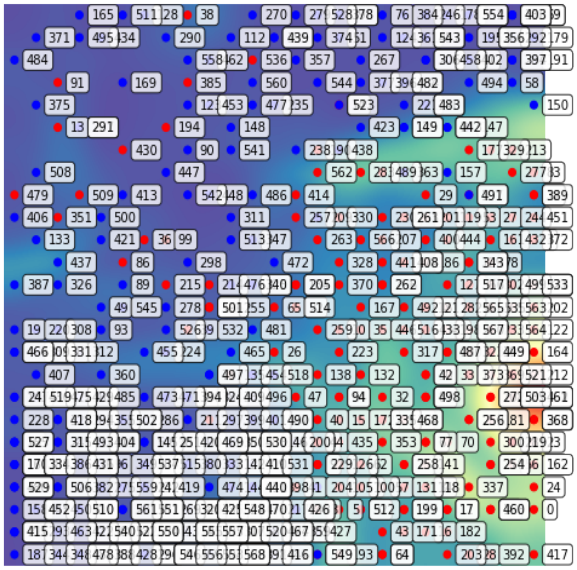
We can see obvious boundary between two clusters. Now we compare these results wtih real data distribution.



It is amazing that there is no big difference between read data distribution and clustering result.It "finds" the boundary between two categories.This is the most interesting result we get.But it still makes mistake when it classifies the data between boundary. This result shows that the clustering algorithm can find the natural structure of dataset if we preserve information of data and the data has obvious category.And this means we can train linear classifier to find boundary for further prediction.

3.6 Self Organizing Map

Now we use SOM on our original dataset and use heat map to show the distance betweeen codebook vectors whose size is 64 and the data distribution in high dimension space(30 dimensions). We create a 8 * 8 topologic map to have visualization result.

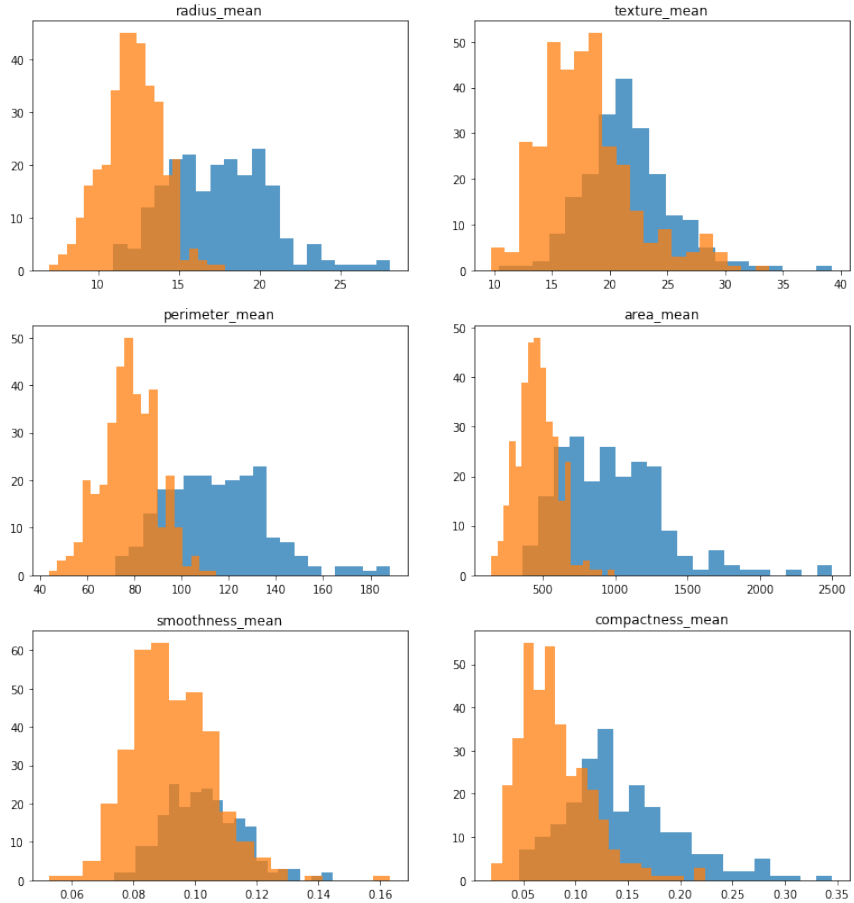


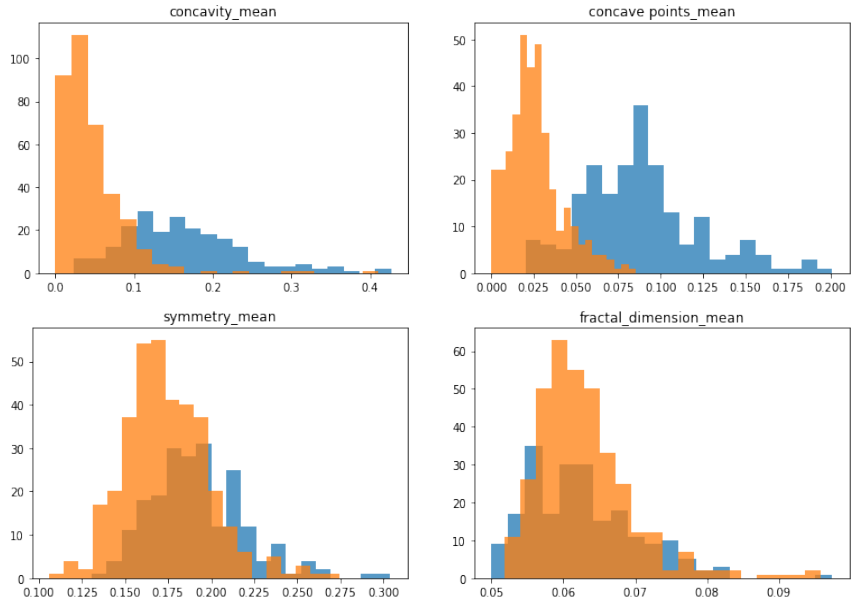
This figure shows that some codebook vectors are closely connected. Others are not so close. The net that is created by codebook vectors is not uniform.But

the most amazing result is that in high dimension space,the data distribution still has boundary but it is not so clear as the clustering result.Both clustering result and SOM find the nature structure of dataset.

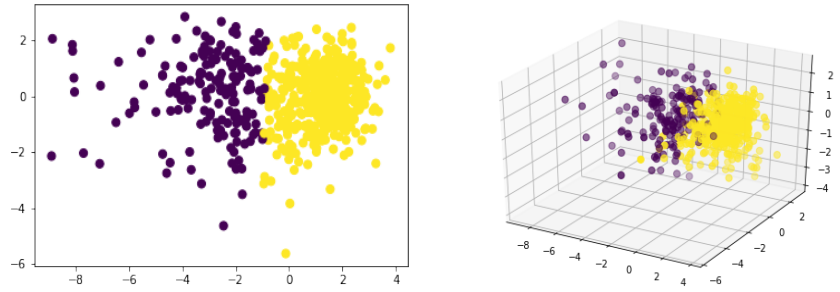
3.7 Final Analysis

Since we find that two categories data can be seperated and the clustering result and topographic map have obvious structure, these shows that there is significant difference between two types of data.Now we try to analyze data on ten feature dimensions.We use histgram to figure out difference between two types of data.(Orange:Benign,Blue:Malignant)





There are obvious differences on each dimension except symmetry mean and fractal dimension mean. For malignant samples, the mean values of each dimension are generally larger than those in benign samples. We think these dimensions preserve the most distinct features.Now we use only these eight dimensions to do PCA and clustering.Here is the clustering result.



This result is closer to read data distribution than using thirty dimensions.We can explain that PCA on 30 dimensions data projects important data on subspace that preserve eight feature dimensions information.That is the reason why the clustering alogrithm works.These figures also explained why the elements in covariance matrix are likely to be greater than 0 as these dimensions are closely related.Beacuse PCA preserves the most important information of dataset in directions of eigenvectors along with the biggest eigenvalues.

4 Conclusion

In this project, we analyze UCI Machine Learning Breast Cancer Wisconsin (Diagnostic) Data Set data by using several algorithms like PCA,SOM and clustering. We successfully prove that PCA can preserve important information in the dataset and elimate useless dimensions. The clustering algorithm can "find out" the natural structure of dataset.This also proves the correctness of clustering algorithm and we can use these algorithms in other dataset.SOM can reflect real data distribution in high dimension space.Our clustering result can be used for further prediction.