भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Identifying fraudulent Taxpayers using Spectral Clustering

Fraud Analytics Assignment 1

Jatin Sharma
CS17BTECH11020

Anjani Kumar
CS17BTECH11002

Dhananjay Raut
CS17BTECH11014

Vijay Tadikamalla
CS17BTECH11040

(Teamsize of 4 with special permission from sir)

Objective

The objective of this assignment is to identify fraudulent taxpayers using clustering techniques on the given dataset. The idea being the algorithm will be able to separate genuine taxpayers from fraudulent ones into separate clusters. Here basic clustering algorithms like K-means will not be that useful as they have a strong tendency to create clusters of nearly equal size and we know that is not the case here. So special clustering algorithm like Spectral clustering comes into rescue as they have a different objective.

Introduction

Tax fraud occurs when an individual or business entity willfully and intentionally falsifies information on a tax return to limit the amount of tax liability. Tax fraud essentially entails cheating on a tax return in an attempt to avoid paying the entire tax obligation. Given the data with various features observed for taxpayers identification for fraudulent taxpayers comes down to an anomaly detection problem. Fraudulent taxpayers will be separate from genuine users and so they are expected to cluster separately by various clustering algorithms. Some algorithms have proven tendency of creating equal-sized clusters and hence will most likely fail to cluster a small size of data points into a separate cluster. But Spectral cluster does not have any such tendency. So we aim to use spectral clustering to separate out possible fraudulent taxpayers. Spectral clustering has its origin in graph theory. It treats each data point as a vertex of the graph and makes use of connectivity information to cluster connected points together. This helps it to cluster a separated set of points into one cluster even though it's size is small.

Dataset

The dataset we got for the assignment consists Of 1163 data points each with 9 real-valued features P1-9. We did not get any feature descriptions and also there was no target column given. It is clearly an unsupervised learning task. We aim to cluster data points in the dataset to possibly identify fraudulent taxpayers which most like be quite separate From other genuine users.

	P1	P2	P3	P4	P5	P6	P7	P8	P9
0	0.595378	-0.531958	0.679654	-0.126799	0.432046	0.988092	-0.029813	0.768742	-0.054167
1	0.982237	0.991481	0.337646	0.228144	0.920032	0.999985	-0.032259	2.161651	-0.054350
2	0.996162	0.893987	0.767413	0.606840	0.970808	0.882602	-0.032267	0.369607	-0.054157
3	0.999928	0.922748	-0.444438	-0.371287	0.528038	-0.221645	-0.032692	-1.065439	0.381914
4	0.985838	0.937512	0.699592	0.585263	0.838804	0.999602	-0.033713	2.089720	-0.054379

Figure 1: Sample data points from the dataset

We did some data analysis on the provided Dataset and its features.

Dataset statistics

Number of variables	9
Number of observations	1163
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	81.9 KiB
Average record size in memory	72.1 B

Figure 2: Various statistics for our dataset

Figure 3 shows the interaction between two sample features in our dataset which clearly shows all the features are not independent. can be effectively represented in smaller dims.

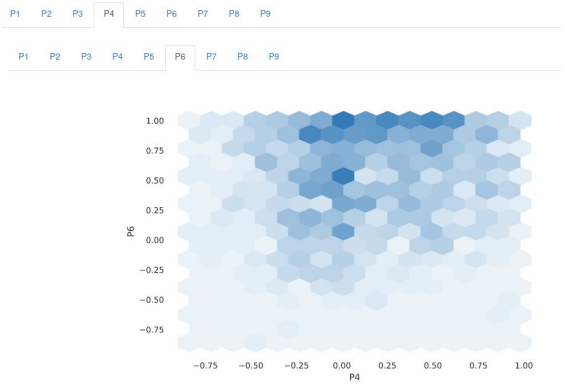


Figure 3: Interaction between P4 and P6.

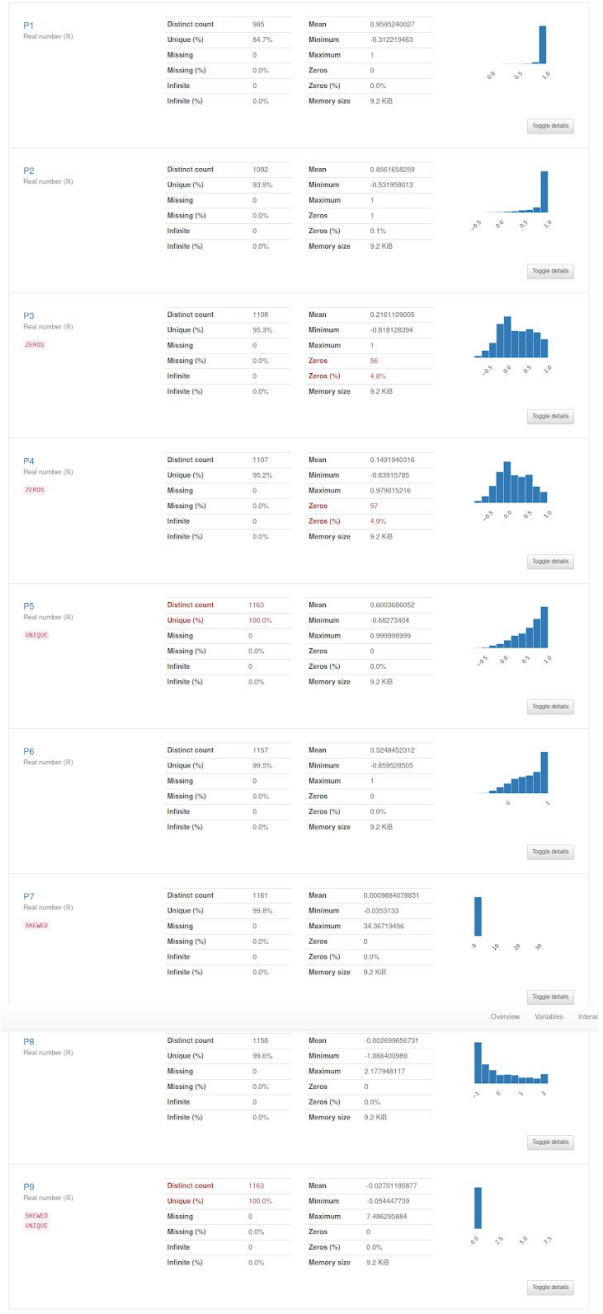


Figure 4: Distribution of all the features of our dataset.

Figure 4 shows various statistics for all the features and their individual distributions. We see some features that are highly skewed. which indicate the need for feature scaling and normalization in our dataset.

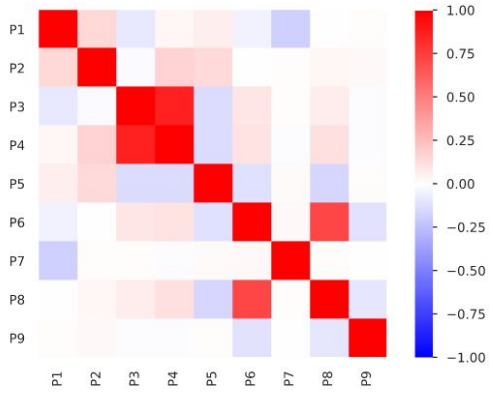


Figure 5: Pearson's r correlation of features.

Figure 5 show Pearson's r correlation which again indicates the presence of lower-dimensional Representation of our dataset.

Now to visualize our data we have to project it in smaller dimensions from current 9 dimensions. So we used the t-SNE technique to get the good visualizable 3-dimensional representation of our dataset. Figure 6 shows The same visualization using Plotly library. This is actually an interactive 3-dimensional plot in the attached notebook.

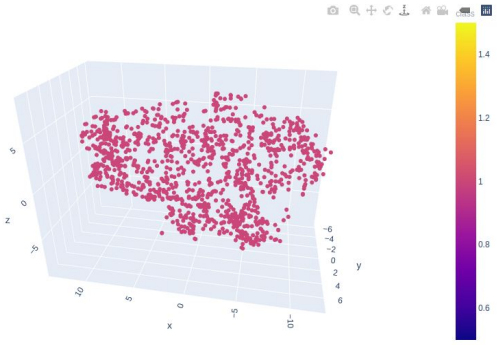


Figure 5: Data points visualized in 3D.

Figure 6 shows why k-means is not a good algorithm in our case. It has clustered our data In nearly equal parts which are not that helpful. In our task to find rare fraudulent taxpayers. And so we use spectral clustering to solve our problem.

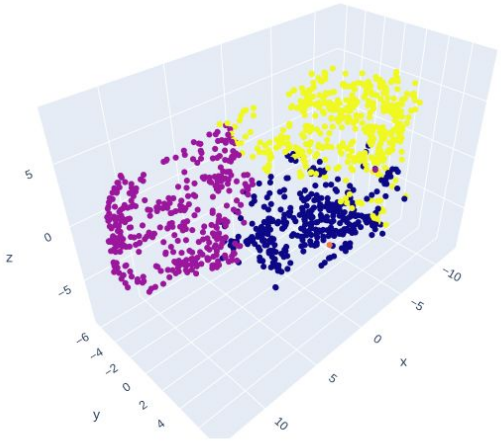


Figure 6: Problem with K-means clustering.

Spectral Clustering

In spectral clustering, data points are treated as nodes of a graph. Thus, spectral clustering is a graph partitioning problem. The nodes are then mapped to a low-dimensional space that can be easily segregated to form clusters. No assumption is made about the shape/form of the clusters. The goal of spectral clustering is to cluster data that is connected but not necessarily compact or clustered within convex boundaries.

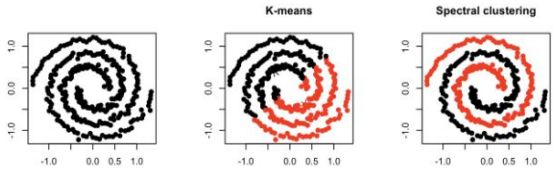


Figure 7: Difference between spectral clustering and K-means clustering.

Algorithm for Spectral Clustering:

- Project data into R^n matrix
- Define an Affinity matrix A.
- Construct the Graph Laplacian from A
- Solve the Eigenvalue problem
- Select k eigenvectors corresponding to the k lowest (or highest) eigenvalues to define a k-dimensional subspace
- Form clusters in this subspace using k-means

We tried two methods for defining Affinity matrix A. first being Gaussian kernel which is defined as follows.

$$A_{ij} = e^{-\alpha \|x_i - x_j\|^2}$$

The second method was using neighbours graph to calculate the affinity.

Along with these, we tried various methods to get Laplacian from Affinity matrix. First, one being the basic method with L = D − A and second method used Normalized Laplacian defined as follows

$$L_N = D^{-1/2} L D^{-1/2}$$

We then solved for eigenvectors of L. and sorted them according to eigenvalues. We chose first k(=2) dimensions for our low dimensional representation.

So we got data points each with smaller 2-dimensional representations. Then we used the standard K-means clustering algorithm on this representation to get clusters. We objectively evaluate the results using t-SNE visualizations with predicted cluster labels to chose best hyperparameter choices.

Results

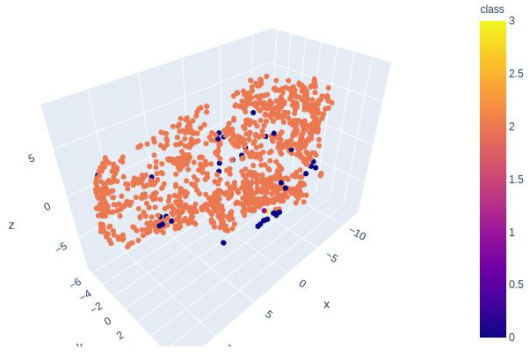


Figure 7: Results after clustering using the basic laplacian with Gaussian affinity matrix.

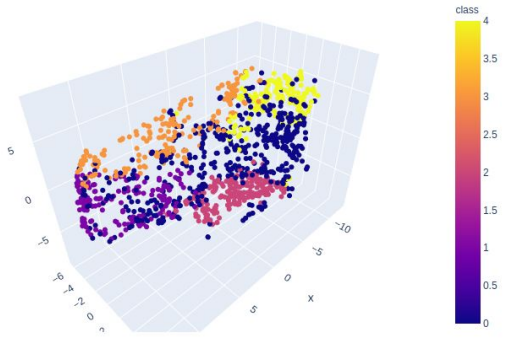


Figure 8: Results after clustering using the normalized laplacian with Gaussian affinity matrix.

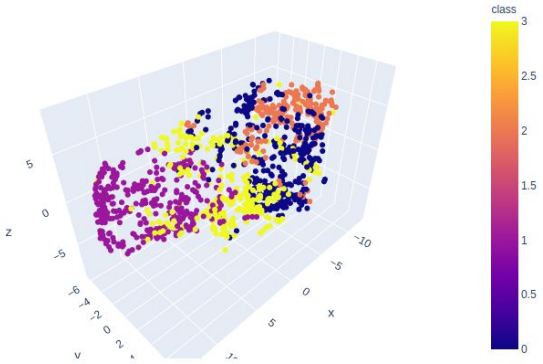
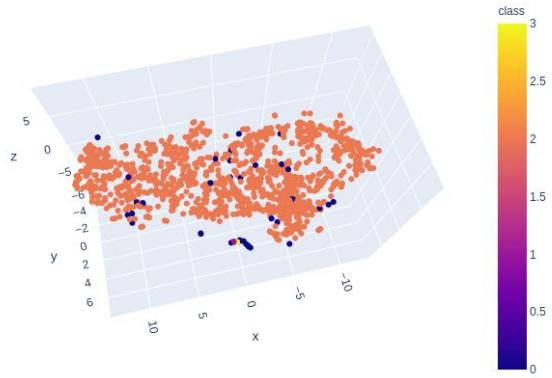


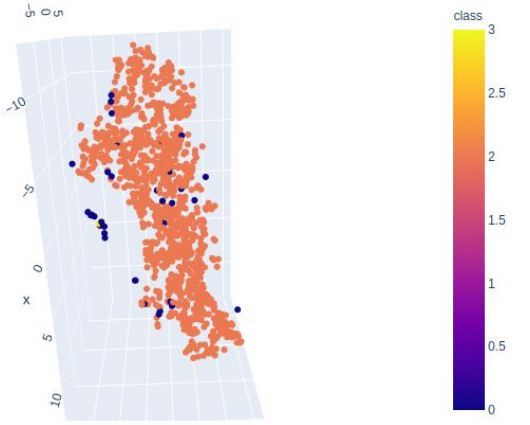
Figure 9: Results after clustering using the basic laplacian with neighbours graph to calculate affinity matrix

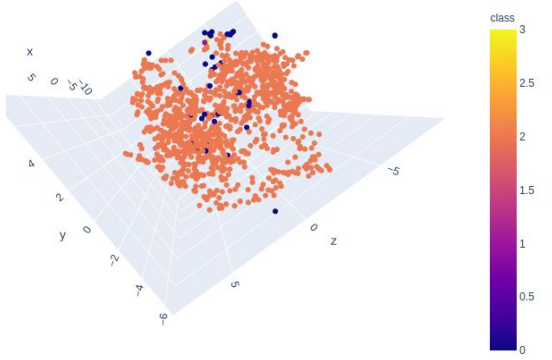
Conclusion

For this particular dataset basic laplacian with a gaussian affinity matrix works well to cluster and separate possible anomalies in our data which can be treated as possible fraudulent taxpayers and should be investigated further.

Visualizations from different Angles:







References

- https://www.kaggle.com/vipulgandhi/spectral-clustering-detailed-explanation
- https://en.wikipedia.org/wiki/Spectral_clustering
- https://towardsdatascience.com/spectral-clustering-aba2640c0d5