K. K. Institute of Engineering Education and Research
Department of Computer Engineering

Project Report

Data Mining and Warehousing

Guided by:
Prof. Dr. S S Sane

By:
Shreyas Kalvankar (17)
Hrushikesh Pandit(18)
Pranav Parwate (19)
Atharva Patil (20)

A.Y. 2020-21 Sem I

Contents

1 Problem Statement

To study, observe and classify the DeepSat-6 Airborne Dataset into 6 categories using Data Mining Algorithms.

1.1 Objectives

- Study the DeepSat-6 Airborne Dataset with respect to application of Data Mining Algorithms.
- Study variations in data using Principle Component Analysis and TSNE.
- Perform feature extraction and data classification by using Convolutional Neural Networks.

2 Introduction

Originally, images were extracted from the National Agriculture Imagery Program (NAIP) dataset. The NAIP dataset consists of a total of 330,000 scenes spanning the whole of the Continental United States (CONUS). The authors used the uncompressed digital Ortho quarter quad tiles (DOQQs) which are GeoTIFF images and the area corresponds to the United States Geological Survey (USGS) topographic quadrangles. The average image tiles are 6000 pixels in width and 7000 pixels in height, measuring around 200 megabytes each. The entire NAIP dataset for CONUS is 65 terabytes. The imagery is acquired at a 1-m ground sample distance (GSD) with a horizontal accuracy that lies within six meters of photo-identifiable ground control points. The images consist of 4 bands - red, green, blue and Near Infrared (NIR).In order to maintain the high variance inherent in the entire NAIP dataset, we sample image patches from a multitude of scenes (a total of 1500 image tiles) covering different landscapes like rural areas, urban areas, densely forested, mountainous terrain, small to large water bodies, agricultural areas, etc. covering the whole state of California.

3 Requirements

3.1 Software Requirements

Sr. no.	**Parameter**	**Requirement**
1	Operating System	Windows
2	IDE	Jupyter Notebook
3	Programming Language	Python3

Table 1: Software Requirements

3.2 Hardware Requirements

Sr. no.	**Parameter**	**Requirement**	**Justification**
1	GPU	Nvidia GeForce GTX 1050	Training the model
2	GPU Memory	>4 GB	Batch Training

Table 2: Hardware Requirements

4 Methodology

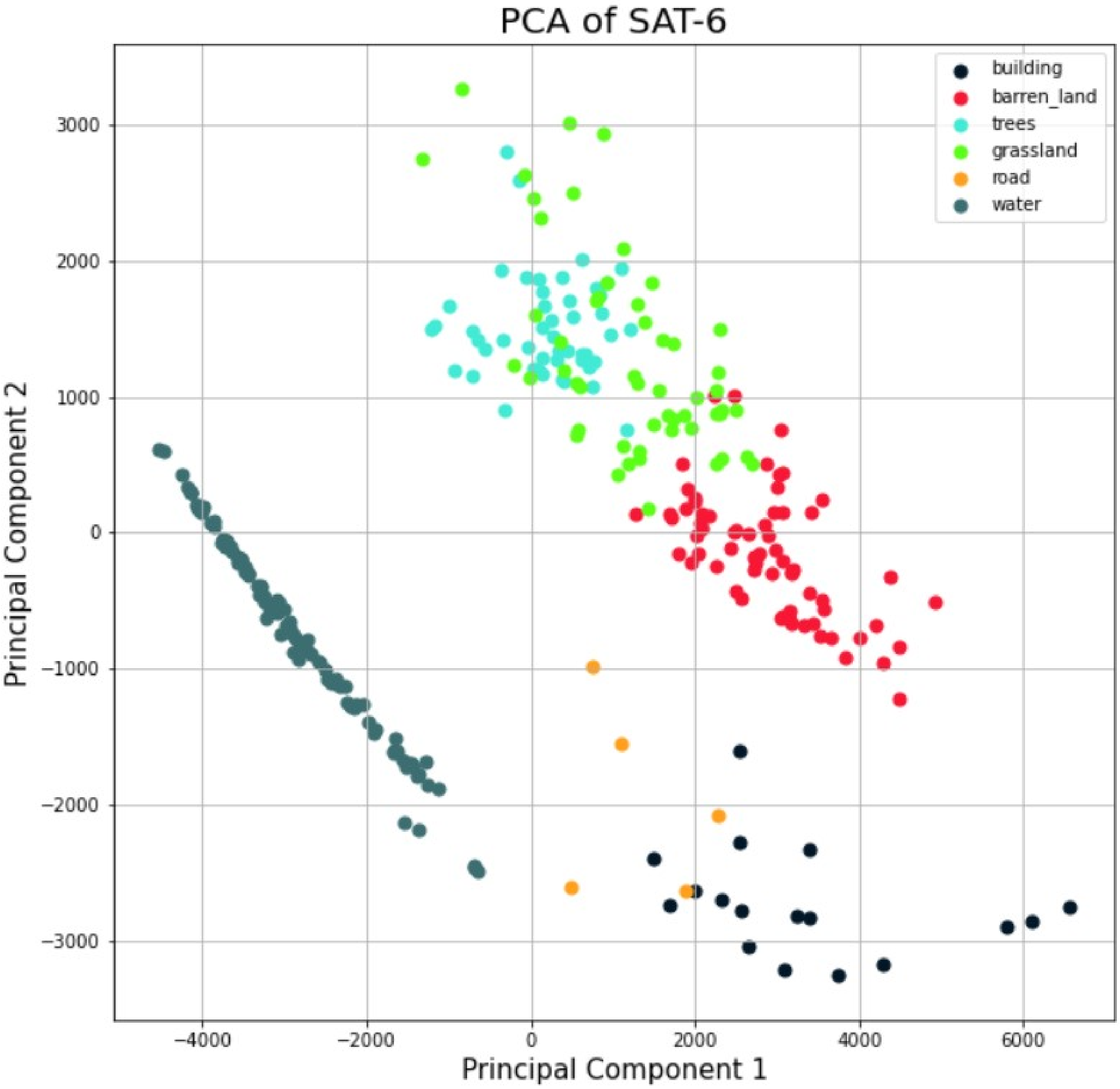
4.1 Algorithm

Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.The principal components of a collection of points in a real p-space are a sequence of p direction vectors, where the i^{th} vector is the direction of a line that best fits the data while being orthogonal to the first $i-1$ vectors. Here, a best-fitting line is defined as one that minimizes the average squared distance from the points to the line.PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. First Component:

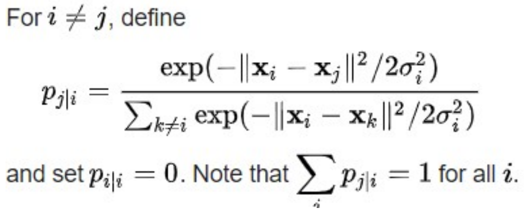
$$\mathbf{w}_{(1)} = \underset{\Vert\mathbf{w}\Vert=1}{\arg\max} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \underset{\Vert\mathbf{w}\Vert=1}{\arg\max} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w}\right)^2 \right\}$$

Further Component :

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^{\mathrm{T}}$$


PCA of SAT-6
Principal Component 1
Principal Component 2
building
barren_land
trees
grassland
road
water
−4000
−2000
0
2000
4000
6000
3000
2000
1000
0
−1000
−2000
−3000

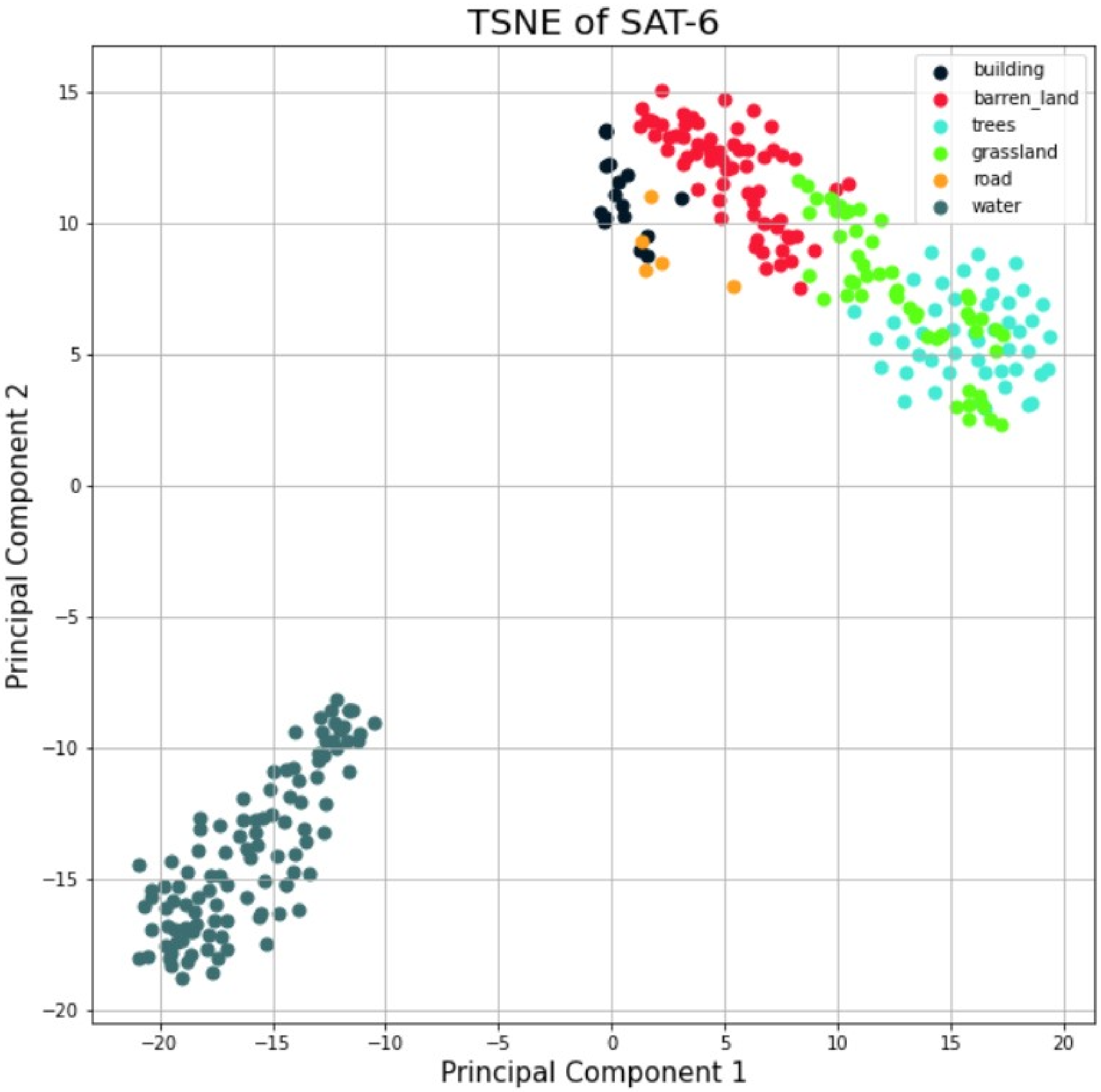
t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map. It is based on Stochastic Neighbor Embedding originally developed by Sam Roweis and Geoffrey Hinton, where Laurens van der Maaten proposed the t-distributed variant.It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.The t-SNE algorithm comprises two main stages. First, t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects are assigned a higher probability while dissimilar points are assigned a lower probability.Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map.



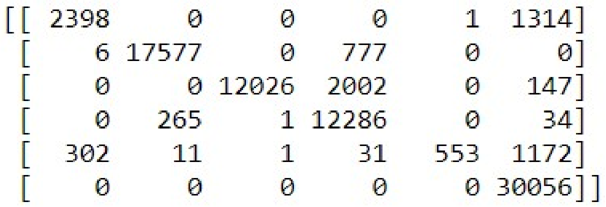
For $i \neq j$, define

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

and set $p_{i|i} = 0$. Note that $\sum_j p_{j|i} = 1$ for all i.


TSNE of SAT-6
building
barren_land
trees
grassland
road
water
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
0
5
10
15
20
Principal Component 1
Principal Component 2

CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The ”fully-connectedness” of these networks makes them prone to overfitting data. Typical ways of regularization include varying the weights as the loss function gets minimized while randomly trimming connectivity. CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in the filters. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme.



```
[[ 2398     0     0     0     1  1314]
 [    6 17577     0   777     0     0]
 [    0     0 12026  2002     0   147]
 [    0   265     1 12286     0    34]
 [  302    11     1    31   553  1172]
 [    0     0     0     0     0 30056]]
```

Figure 1: Confusion Matrix

5 Results

	precision	recall	f1-score	support
building	0.89	0.65	0.75	3713
barren_land	0.98	0.96	0.97	18360
trees	1.00	0.85	0.92	14175
grassland	0.81	0.98	0.89	12586
road	1.00	0.27	0.42	2070
water	0.92	1.00	0.96	30056
accuracy	—	—	0.93	80960
macro avg	0.93	0.78	0.82	80960
weighted avg	0.93	0.93	0.92	80960

Table 3: Classification Report

6 Conclusion

Thus we have successfully mined the DeepSat - 6 Airborne Dataset using Principle Component Analysis and t-distributed stochastic neighbor embedding (t-SNE).