Sources of Big Data

by

S.Lakshmivarahan
School of Computer Science
University of Oklahoma
Norman, OK-73019, USA
varahan@ou.edu

Regime change - data sparse to data rich

- Early days of Astronomy: Data collected by humans using telescopes
- Early days of Meteorology: Baloons, Ships, Aircrafts

Advances in Technology

- Sensors
- Wireless Communication
- Large scale storage devices
- Computer with ever increasing power - Tera, Peta Flops
- Data collected doubles in 2-3 years

Sources / Examples of Big data

- Speech signals
- Radar signals
- Hyperspectral images from satellites
- Genome analysis

- Text documents
- Finger prints
- Facial recognition
- Climate data

- Impressions of friction ridges of part or all of a human finger
- Assume a resolution of $m = 64 \times 64 = 4096 = 2^{12}$ pieces of information
- FBI may have n (millions) such images
- Data matrix: $X \in R^{m \times n}$

- A color image is on a 256×256 grid with three colors - R, B, G
- The value of $m = 256 \times 256 \times 3 = 2^{16} \times 3$ $= 3 \times 65,536$
- An FBI data base may contain n millions of such images
- For each face, one may have 10 variations depending illumination condition, different poses, facial expressions, time in a day etc.,

Hyper Spectral Satellite Images

- Used for Geological / Geographical scenes
- Use wavelength: 350 nm - 3,500 nm range of the electromagnetic spectrum including the visible and infrared
- Visible range - 380 nm - 700 nm, Infrared - 700 nm - 3500 nm, Microwave - 1 mm - 1 meter
- Range is divided into small spectral bands of width 5 - 10 nm.
- Each spectral band creates an image of the scene

Hyper Spectral Image (HSI)

- Image is a 3-D object - HSI cube
- Image information from a given spectral band reside in a 2D plane
- For a given pixel, the vertical dimension provides the spectral radiance
- f(x, y, s) $a \leq x \leq b$, $c \leq y \leq d$, $s_1 \leq s \leq s_2$

- JPL, California Institute of Technology
  https://aviris.jpl.nasa.gov/data/get_aviris_data.html
- L. Biehl: https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html

Text documents

- Assume that there are m-words in the present dictionary
- Given a text, create an m vector where the i^{th} element is the frequency of occurrence of the i^{th} word of the dictionary in that text
- Each text is represented by a vector of frequencies of occurrence of words
- Data matrix $\bar{X} \in R^{m \times n}$ represents the collection of n texts

Text document sources

- sci.crypt sci.space
- sci.med sci.religion.christian

- Atmosphere, ocean, pure water mass, deserts
- A typical field variable: f(x,y,z,t)
- Radius of earth: R= 4,000 miles
- Equatorial circumference: $2\pi R = 25,000$ miles
- Surface area: $4\pi R^2 \approx 469 \times 10^6$ squaremiles

Size of Grid

- Assume a grid of area 1 square mile
- Need nearly 500×10^6 grids
- Count 50 levels, covering say 10 miles above sea level
- Total number of grid points (x, y, z) $\approx 25 \times 10^9$
- Time intervals, t: Day, month, year etc

Radar network in the USA

- Radio Detection and Ranging
- NWS operates Doppler (WSR-88D) radars: 142 in the lower 48 states + 11 in Hawaii and Alaska
- National Center for Environmental Information (NCEI) archieves the radar data and Terminal Doppler radars

Radar frequency bands

Band name	Frequency range	Wavelength range	Notes
HF	3–30 MHz	10–100 m	Coastal radar systems, over-the-horizon radar (OTH) radars; 'high frequency'
VHF	30–300 MHz	1–10 m	Very long range, ground penetrating; 'very high frequency'
P	< 300 MHz	> 1 m	'P' for 'previous', applied retrospectively to early radar systems; essentially HF + VHF
UHF	300–1000 MHz	0.3–1 m	Very long range (e.g. ballistic missile early warning), ground penetrating, foliage penetrating; 'ultra high frequency'
L	1–2 GHz	15–30 cm	Long range air traffic control and surveillance; 'L' for 'long'
S	2–4 GHz	7.5–15 cm	Moderate range surveillance, Terminal air traffic control, long-range weather, marine radar; 'S' for 'short'
C	4–8 GHz	3.75–7.5 cm	Satellite transponders; a compromise (hence 'C') between X and S bands; weather; long range tracking
X	8–12 GHz	2.5–3.75 cm	Missile guidance, marine radar, weather, medium-resolution mapping and ground surveillance; in the United States the narrow range 10.525 GHz ±25 MHz is used for airport radar; short range tracking. Named X band because the frequency was a secret during WW2.
K_u	12–18 GHz	1.67–2.5 cm	High-resolution, also used for satellite transponders, frequency under K band (hence 'u')

Band name	Frequency range	Wavelength range	Notes
K	18–24 GHz	1.11–1.67 cm	From German *kurz*, meaning 'short'; limited use due to absorption by water vapour, so K_u and K_a were used instead for surveillance. K-band is used for detecting clouds by meteorologists, and by police for detecting speeding motorists. K-band radar guns operate at 24.150 ± 0.100 GHz.
K_a	24–40 GHz	0.75–1.11 cm	Mapping, short range, airport surveillance; frequency just above K band (hence 'a') Photo radar, used to trigger cameras which take pictures of license plates of cars running red lights, operates at 34.300 ± 0.100 GHz.
mm	40–300 GHz	1.0–7.5 mm	Millimetre band, subdivided as below. The frequency ranges depend on waveguide size. Multiple letters are assigned to these bands by different groups. These are from Baytron, a now defunct company that made test equipment.
V	40–75 GHz	4.0–7.5 mm	Very strongly absorbed by atmospheric oxygen, which resonates at 60 GHz.
W	75–110 GHz	2.7–4.0 mm	Used as a visual sensor for experimental autonomous vehicles, high-resolution meteorological observation, and imaging.

Radar modulators[edit]

Satellite data

- Russia launched the first satellite- Sputnik in 1957
- Weather satellite was launched in 1960
- There are about 1,100 active satellites
- These can be geostationary, Polar orbiting
- Meteorological, GPS, Military

Problems in Data Analytics

- Classification / clustering of data
- Processing queries in large data base
- Feature Extraction / Dimensionality Reduction
- Solution of large scale least square problems

Algorithms and Complexity : Illustration

- Given a data matrix $\bar{X} \in R^{m \times n}$, m>n
- Let $X = \frac{1}{\sqrt{N}} \bar{X}$
- Data may be Correlated : Data = Signal + Noise
- Want to extract the signal and express it as a linear combination of a small number of uncorrelated components
- This is the basis for Principal Component Analysis (PCA)

- Compute the Gramian matrix: $X^T X \in R^{n \times n}$
- Compute the eigen decomposition of $X^T X$:

$$(X^T X)V = V\Lambda$$

- $V = [V_1, V_2, \ldots, V_n] \in R^{n \times n}$
- $\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_n) \in R^{n \times n}$
- $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$

Algorithm

- Compute $U = [U_1, U_2, \ldots, U_n] \in R^{m \times n}$
- $U_i = \frac{1}{\sqrt{\lambda_i}} X V_i \in R^m$
- Singular value decomposition of X:

$$X = U \Lambda^{1/2} V^T$$

Algorithm

- Let $0 < \beta < 1$ be given
- Find k: $\sum_{i=1}^{k} \lambda_i \geq (1-\beta) \sum_{i=1}^{n} \lambda_i$
- Define $\bar{U} = [U_1, U_2, \ldots, U_k] \in R^{m \times k}$
  $\bar{V} = [V_1, V_2, \ldots, V_k] \in R^{n \times k}$
  $\bar{\Lambda} = Diag(\lambda_1, \lambda_2, \ldots, \lambda_k) \in R^{k \times k}$

Algorithm

- Then
$$\bar{X} \approx \bar{U}\bar{\Lambda}^{-1/2}\bar{V}^T$$
- $Y = \bar{U}^T\bar{X} = \bar{\Lambda}^{1/2}\bar{V} \in R^{k\times n}$
- Y is the compressed version of X
- Example of dimension reduction from m to k

Computational Complexity

- Cost of $X^T X$: $O(mn^2)$
- Cost of finding V, Λ: $O(n^3)$
- Total cost: $O(n^2(n+m))$

Example:

- Consider: $n = 10^6$, $n^3 = 10^{18}$ operations
- Time/operation: 10^{-12} sec - Tera Flop m/c
- Total time: $\frac{10^{18}}{10^{12}} = 10^6$ sec

Example:

- Number of seconds in a day $= 60 \times 60 \times 24 = 86,400 = 0.864 \times 10^5$
- Number of days $= \frac{10^6}{.864 \times 10^5} = \frac{10}{.864} = 11.58$ days

Summary

- We are moving from data sparse to data rich regime
- More data is not merrier
- They may be correlated - provide less information
- More data implies larger computational time

Summary

- This calls for techniques to reduce the data set - data reduction techniques
- PCA is optimal, but data dependent
- There is a growing need for data independent tools that can work across the spectrum of datasets from various domain