CS686 DATA MINING

EEG BRAINWAVE FOR CONFUSION

MIN CHEN THANAWUT ANANPIRIYAKUL SHIYI TAN

AGENDA

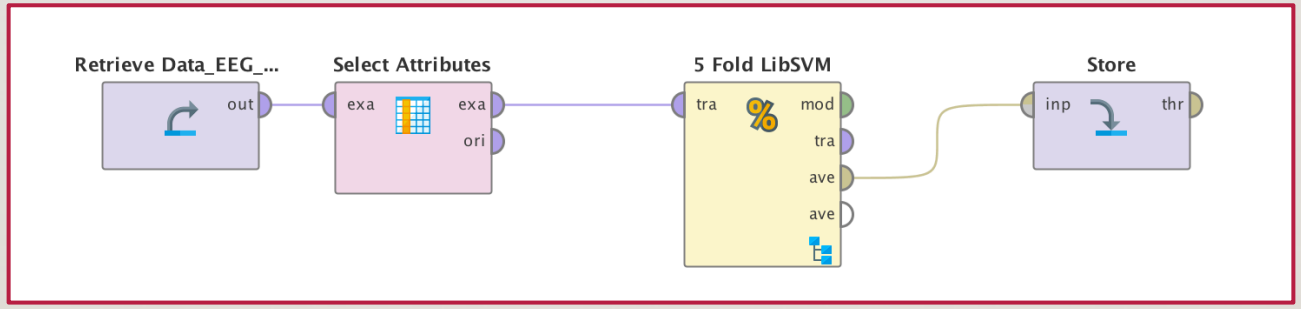
- Experiment & Dataset
- Tools and Frameworks
- Problem Description
- Data Preprocessing
- Models and Evaluation
- Challenges

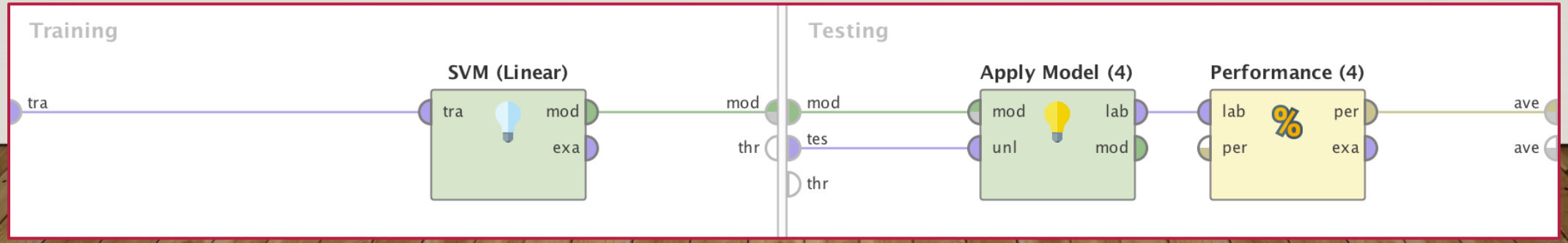
EXPERIMENT & DATASET

- EEG: An electrophysiological monitoring method to record **electrical activity of the brain.**
- EEG data were collected from **10 college students** watching at **10 videos**.
- Each video can be confusing or not confusing. Students rated them during the experiment.
- **1-channel wireless headset** was used to record brainwave.
- We have **12811 rows** and **15 columns**.
- The data was aggregated. **1 row means 0.5 second.**

TOOLS & FRAMEWORKS

- **Rapidminer** (Good for Exploratory Analysis)
  - Powerful & Easy to Use
  - Drag & Drop (with a little coding)
  - Visualize data and Get Threshold of Models' Accuracy
- **R Programming Language**
  - Powerful & Flexible
  - Need Coding
  - Take Time to Develop





PROBLEM DESCRIPTION

- **We want to predict confusion of lecture videos from EEG data.**
- Problem Type: **Binary Classification**
- Target: **Confusing** (Yes / No)
- Features:
  - Attention
  - Meditation
  - Raw
  - Delta
  - Theta
  - Alpha1
  - Alpha2
  - Beta1
  - Beta2
  - Gamma1
  - Gamma2
  - Subject ID
  - Video ID

DATA PREPROCESSING

- Feature Selection
- Normalization
- Feature Generation
- Data Aggregation

FEATURE SELECTION

- No strong relationship has been found between each feature and target.
- The highest correlation is 0.15 and the lowest correlation is -0.12.

	Attention	Meditation	Raw	Delta	Theta	Alpha1	Alpha2	Beta1	Beta2	Gamma1	Gamma2	SelfDefinedConfusion
Attention	1.00000000	0.54860617	-0.094932609	-0.22040251	-0.2583561	-0.19866152	-0.24701101	-0.25536124	-0.30662414	-0.29990260	-0.31273425	-0.119087999
Meditation	0.54860617	1.00000000	-0.106923088	-0.11181024	-0.1798871	-0.04327509	-0.15373968	-0.24077411	-0.39136526	-0.34492192	-0.37325275	-0.023502268
Raw	-0.09493261	-0.10692309	1.000000000	0.04890827	0.0534425	0.04524732	0.06386737	0.09086331	0.17363074	0.18688406	0.20577189	-0.001464805
Delta	-0.22040251	-0.11181024	0.048908272	1.00000000	0.4833865	0.42652984	0.40989774	0.39521675	0.26081956	0.21006246	0.21867276	0.148588296
Theta	-0.25835609	-0.17988708	0.053442501	0.48338648	1.0000000	0.58695601	0.55379594	0.55554088	0.36897517	0.30046781	0.30857500	0.147557833
Alpha1	-0.19866152	-0.04327509	0.045247318	0.42652984	0.5869560	1.00000000	0.56523189	0.54911327	0.31792262	0.26767665	0.32359258	0.121711328
Alpha2	-0.24701101	-0.15373968	0.063867374	0.40989774	0.5537959	0.56523189	1.00000000	0.62682443	0.46483995	0.37497882	0.45556704	0.107238239
Beta1	-0.25536124	-0.24077411	0.090863310	0.39521675	0.5555409	0.54911327	0.62682443	1.00000000	0.48049006	0.42216482	0.52748396	0.111481557
Beta2	-0.30662414	-0.39136526	0.173630736	0.26081956	0.3689752	0.31792262	0.46483995	0.48049006	1.00000000	0.80893045	0.69258415	0.019555883
Gamma1	-0.29990260	-0.34492192	0.186884063	0.21006246	0.3004678	0.26767665	0.37497882	0.42216482	0.80893045	1.00000000	0.73983649	0.011316171
Gamma2	-0.31273425	-0.37325275	0.205771893	0.21867276	0.3085750	0.32359258	0.45556704	0.52748396	0.69258415	0.73983649	1.00000000	0.053532989
SelfDefinedConfusion	-0.11908800	-0.02350227	-0.001464805	0.14858830	0.1475578	0.12171133	0.10723824	0.11148156	0.01955588	0.01131617	0.05353299	1.000000000

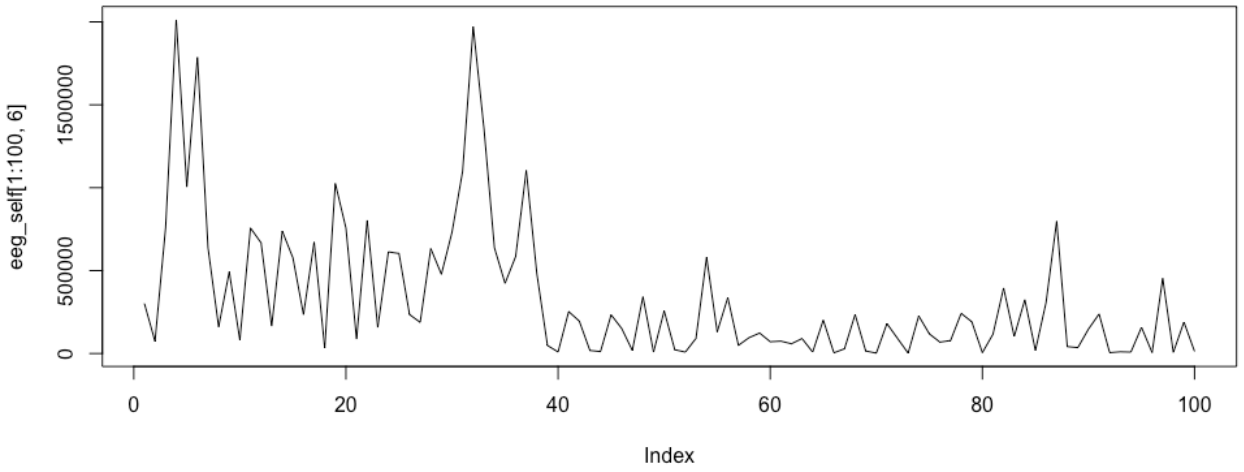
FEATURE SELECTION

- We only have 13 features.
- Only **Subject ID and Video ID have been removed.**
  - Attention
  - Meditation
  - Raw
  - Delta
  - Theta
  - Alpha1
  - Alpha2
  - Beta1
  - Beta2
  - Gamma1
  - Gamma2
  - ~~Subject ID~~
  - ~~Video ID~~
- **Removing more features does not help** increase the accuracy.

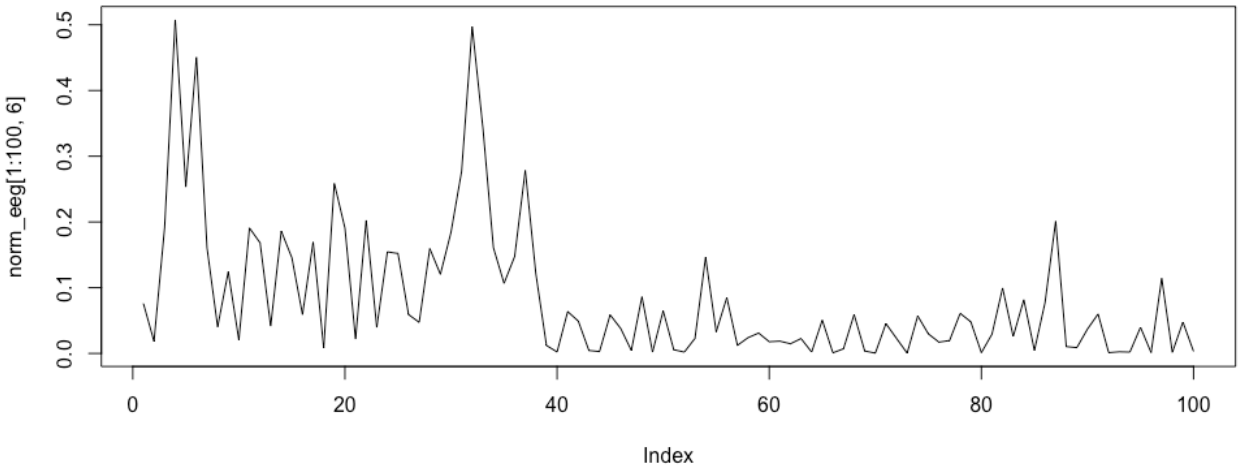
NORMALIZATION

- Each feature has different range.
- Range can be in 100 or up to a few 10000s.
- Use "preProcess" function from "caret" package.
  - **Method "range" will scale data into range [0,1].**

Plot of Delta before Normalization



Plot of Delta after Normalization



FEATURE GENERATION

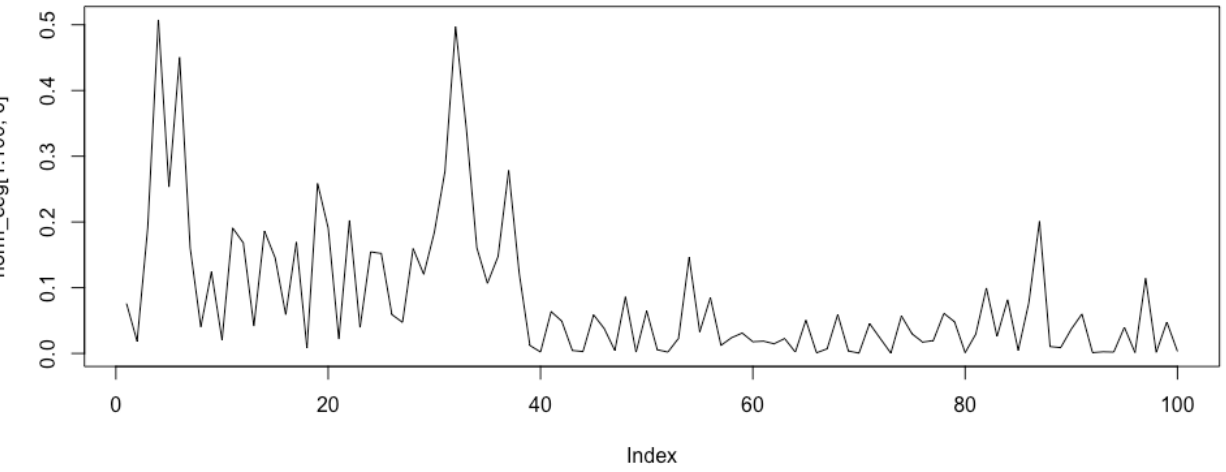
- Since our data are time series.
- **Differences of data in each row are added as features.**
- Now we have 22 features.

Attention	ΔAttention	Meditation	ΔMeditation	Raw	ΔRaw
0.68		0.5		0.8	
0.61	-0.07	0.52	+0.02	0.67	-0.13
0.63	+0.02	0.51	-0.01	0.72	+0.05

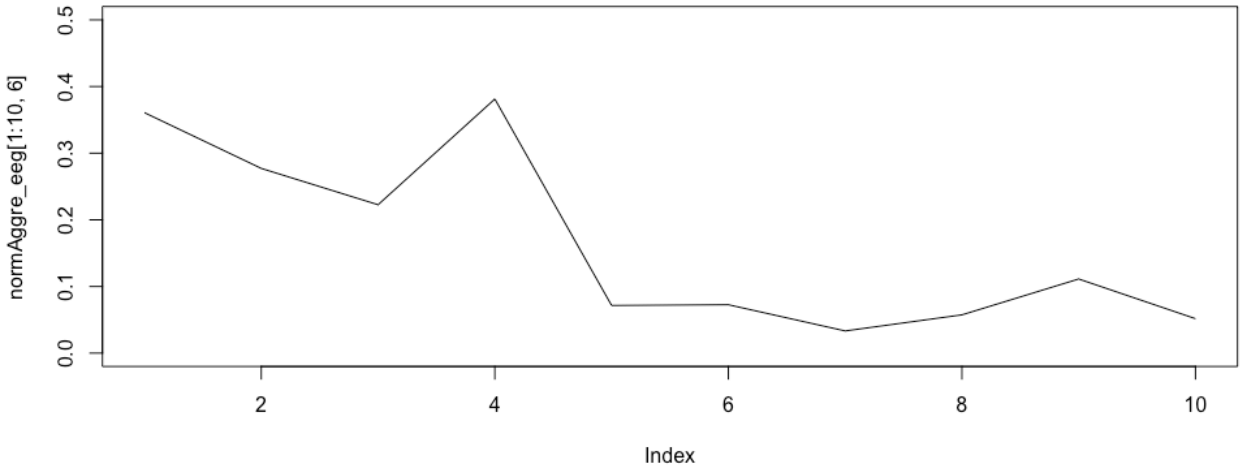
DATA AGGREGATION

- Aggregate data from 0.5 second to 5 seconds.
- This method will help smoothen the data.
- Our goal is to predict confusion of a video (2 mins) not just a data point (0.5 second).

Plot of Delta before Aggregation



Plot of Delta after Aggregation

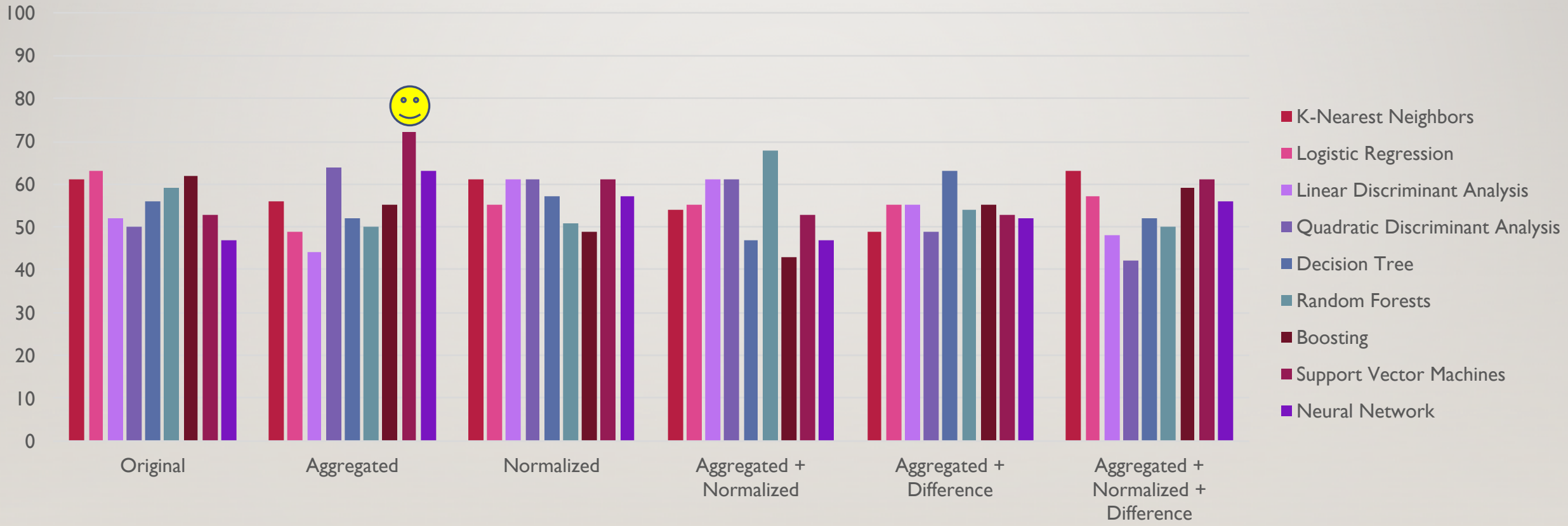


MODEL AND EVALUATION

- Selected Models
  - K-Nearest Neighbors
  - Logistic Regression
  - Linear Discriminant Analysis
  - Quadratic Discriminant Analysis
  - Decision Tree
  - Random Forests
  - Boosting
  - Support Vector Machines
  - Neural Network

- We have 10 subjects and 10 videos.
- We have 100 cases in total.
- We do Leave-One-Out Cross Validation.

ACCURACY


100
90
80
70
60
50
40
30
20
10
0
Original
Aggregated
Normalized
Aggregated + Normalized
Aggregated + Difference
Aggregated + Normalized + Difference
K-Nearest Neighbors
Logistic Regression
Linear Discriminant Analysis
Quadratic Discriminant Analysis
Decision Tree
Random Forests
Boosting
Support Vector Machines
Neural Network

CHALLENGES

- The data is only one channel.
  - 1-channel wireless headset was used in the experiment.
- The data is very noisy.
  - The electrode was placed at the skin.