Visual Analysis of Submodular Point and Feature Selection for Data-Efficient Machine Learning

Outline of the Talk

- Motivation
- Contributions
- Our Framework
- Datasets
- Results
- Conclusions

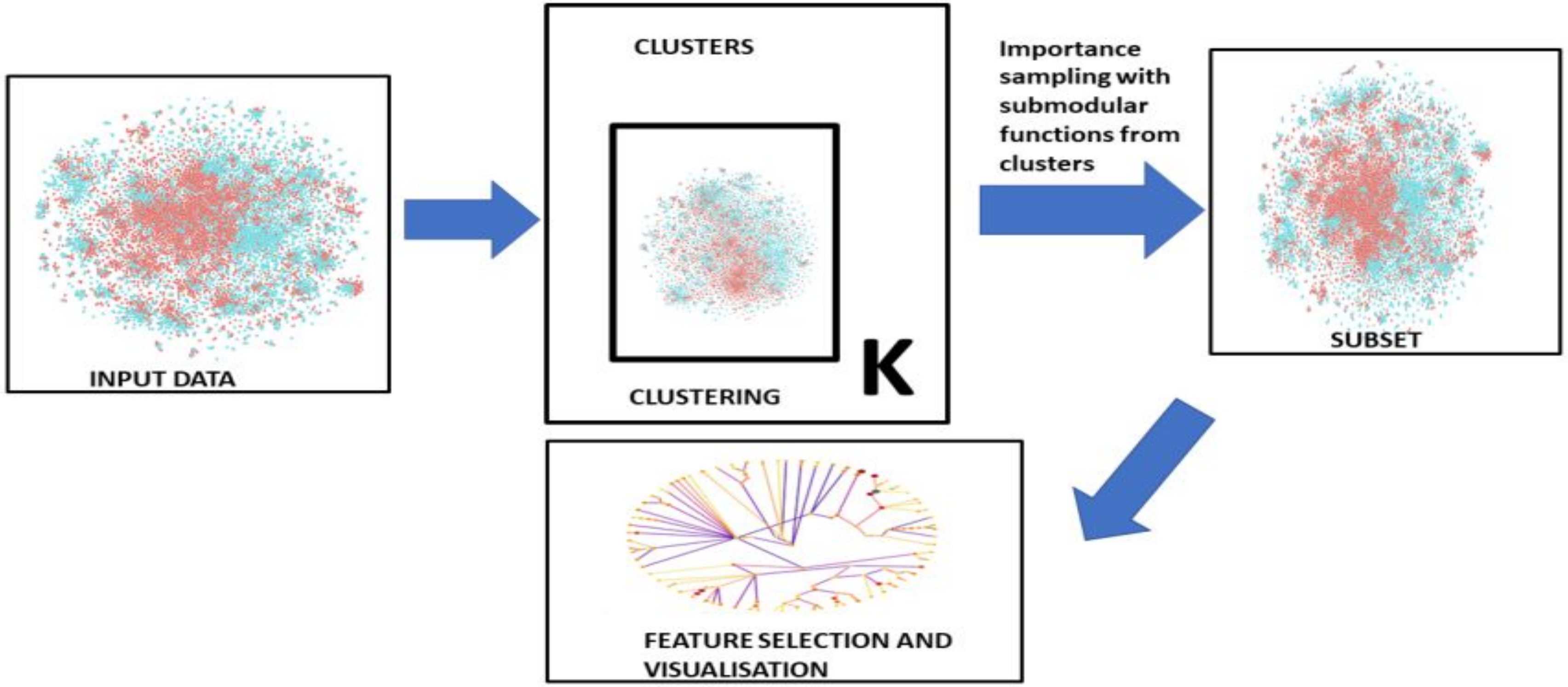
Motivation

- Lots of data are required to train machine learning models
- Theoretically the more the data, the better the performance
- More data means more compute times required
- Data also normally contains redundant and irrelevant data points
- **This projects aims to train machine learning models by selecting the most important important points without sacrificing performance**
- We achieve this with submodular data selection combined with clustering

Our contributions

1. We propose a two-stage subset selection algorithm that combines clustering and importance sampling to induce diversity with subsets and reduce computational requirements
2. We perform experimental evaluation of different point selection methods and their effect on the performance of machine learning models
3. We demonstrate the effect of sampling method on the lower dimensional spaces using multi-dimensional projections

Our Framework


CLUSTERS
Importance sampling with submodular functions from clusters
INPUT DATA
K
CLUSTERING
SUBSET
FEATURE SELECTION AND VISUALISATION

Datasets

IMDB Dataset (Text)

- Large Movie Review Dataset. This is a dataset for binary sentiment classification
- Dataset contains a set of 25,000 highly polar movie reviews for training, and 25,000 for testing

Fashion MNIST (Image)

- Fashion-MNIST is a dataset of Zalando's article images
- Dataset consists of a training set of 60,000 examples and a test set of 10,000 examples.
- Each example is a 28x28 grayscale image, associated with a label from 10 classes.

Models Used for classification

IMDB (Text Classification)

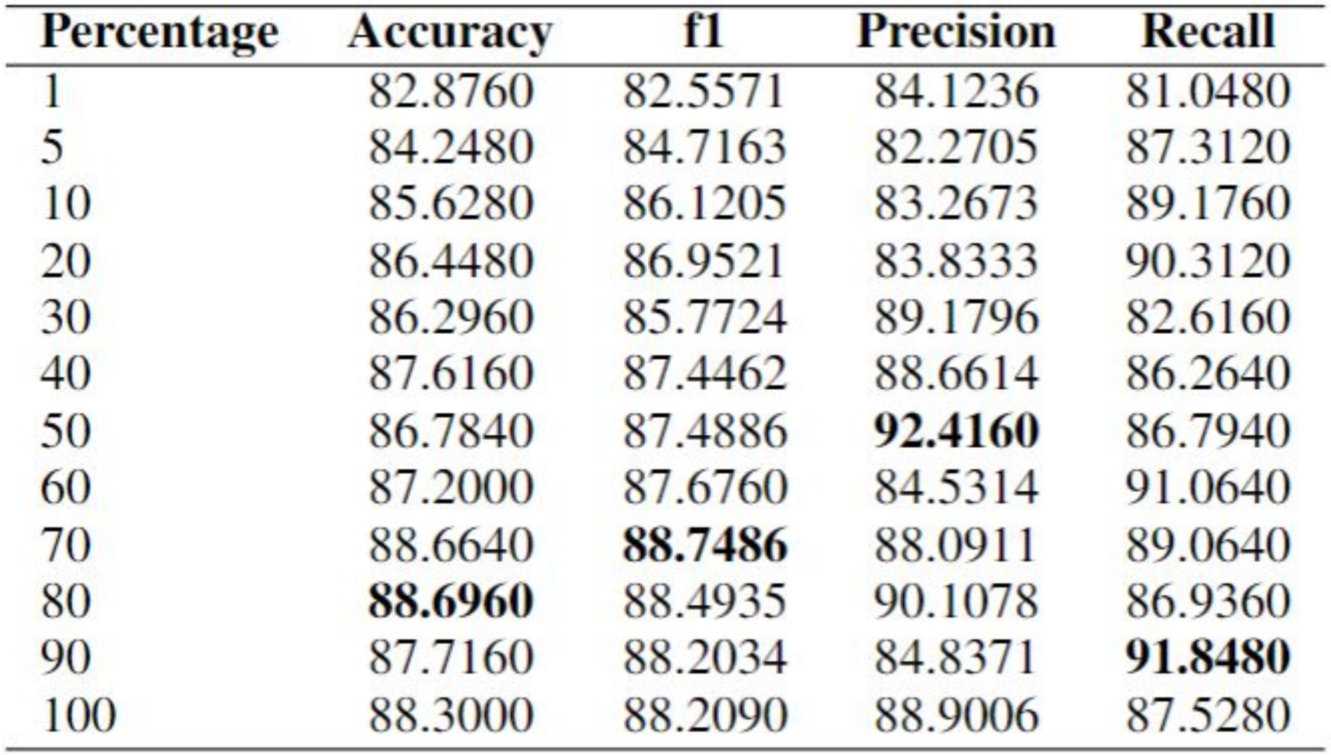
Used RoBERTa (Robustly Optimised BERT)

Fashion MNIST (Image Classification)

Used CNN (Convolutional Neural Network)

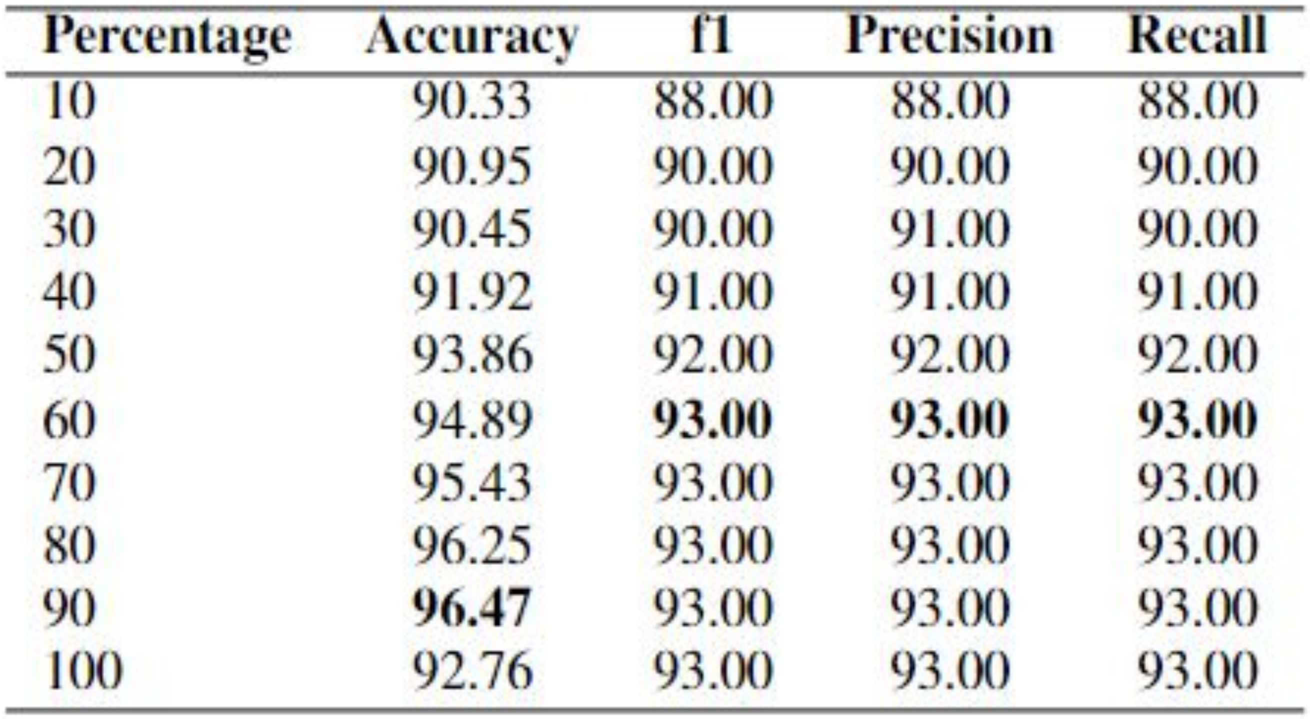
RESULTS AND DISCUSSION

Experiment Results on IMDB Dataset



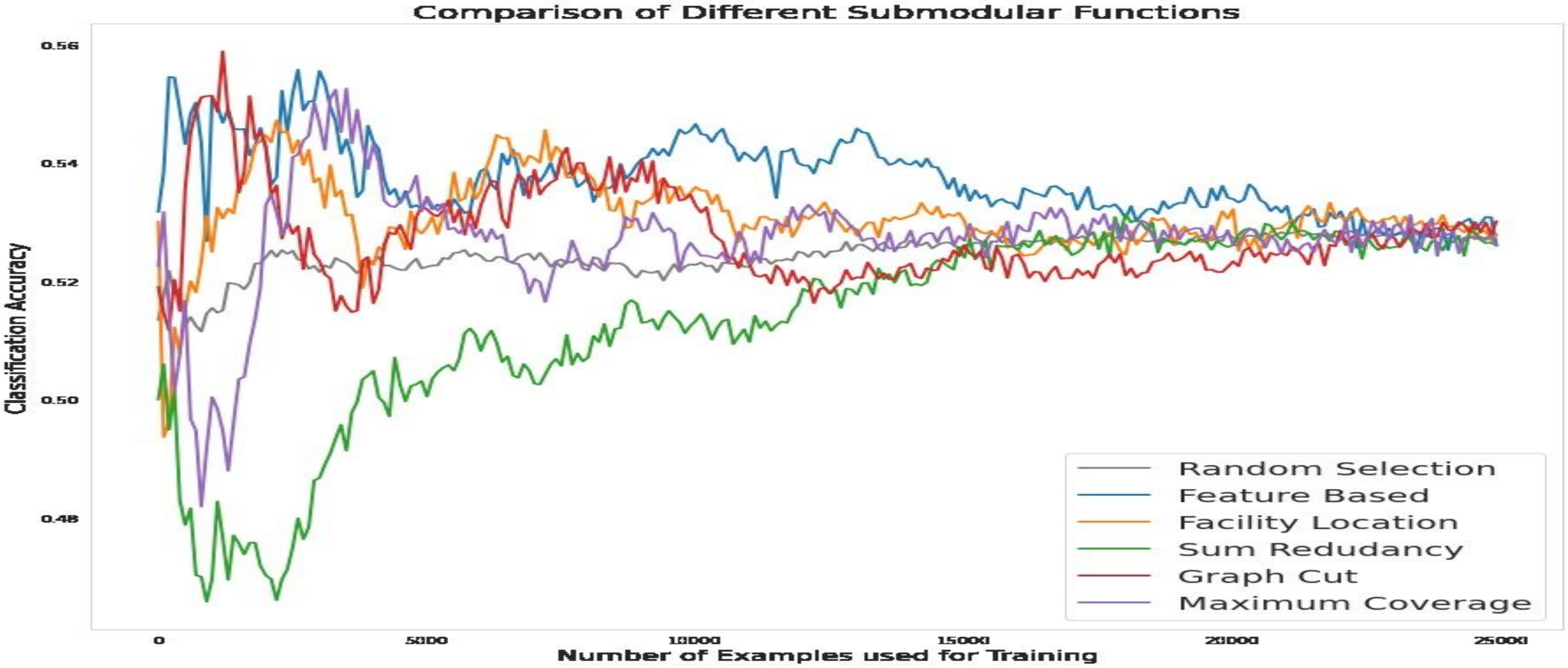
Percentage	Accuracy	f1	Precision	Recall
1	82.8760	82.5571	84.1236	81.0480
5	84.2480	84.7163	82.2705	87.3120
10	85.6280	86.1205	83.2673	89.1760
20	86.4480	86.9521	83.8333	90.3120
30	86.2960	85.7724	89.1796	82.6160
40	87.6160	87.4462	88.6614	86.2640
50	86.7840	87.4886	**92.4160**	86.7940
60	87.2000	87.6760	84.5314	91.0640
70	88.6640	**88.7486**	88.0911	89.0640
80	**88.6960**	88.4935	90.1078	86.9360
90	87.7160	88.2034	84.8371	**91.8480**
100	88.3000	88.2090	88.9006	87.5280

Experimental Results on Fashion MNIST

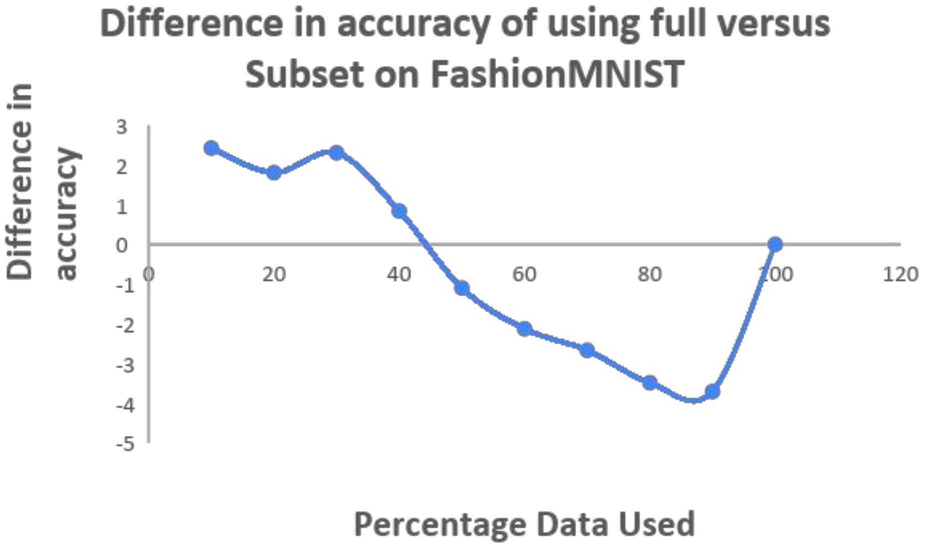


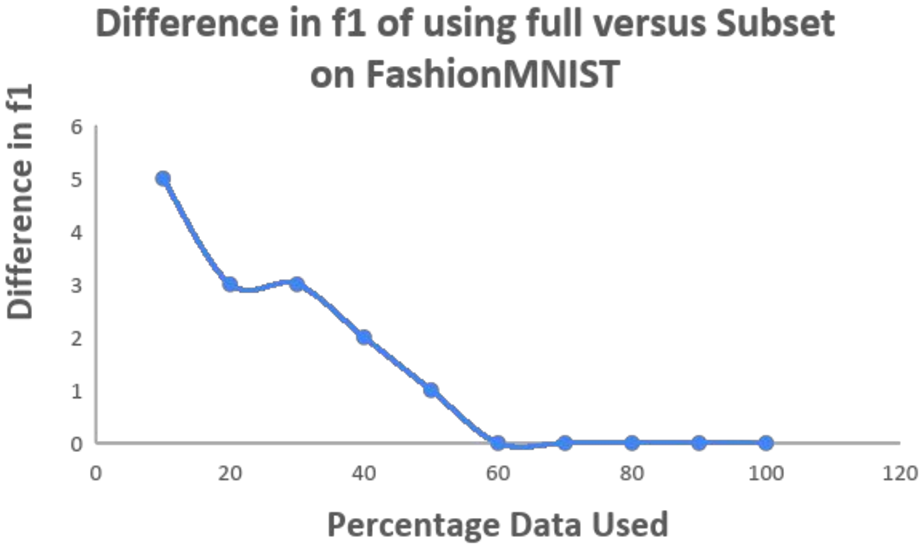
Percentage	Accuracy	f1	Precision	Recall
10	90.33	88.00	88.00	88.00
20	90.95	90.00	90.00	90.00
30	90.45	90.00	91.00	90.00
40	91.92	91.00	91.00	91.00
50	93.86	92.00	92.00	92.00
60	94.89	**93.00**	**93.00**	**93.00**
70	95.43	93.00	93.00	93.00
80	96.25	93.00	93.00	93.00
90	**96.47**	93.00	93.00	93.00
100	92.76	93.00	93.00	93.00

Does Selection Method Matter

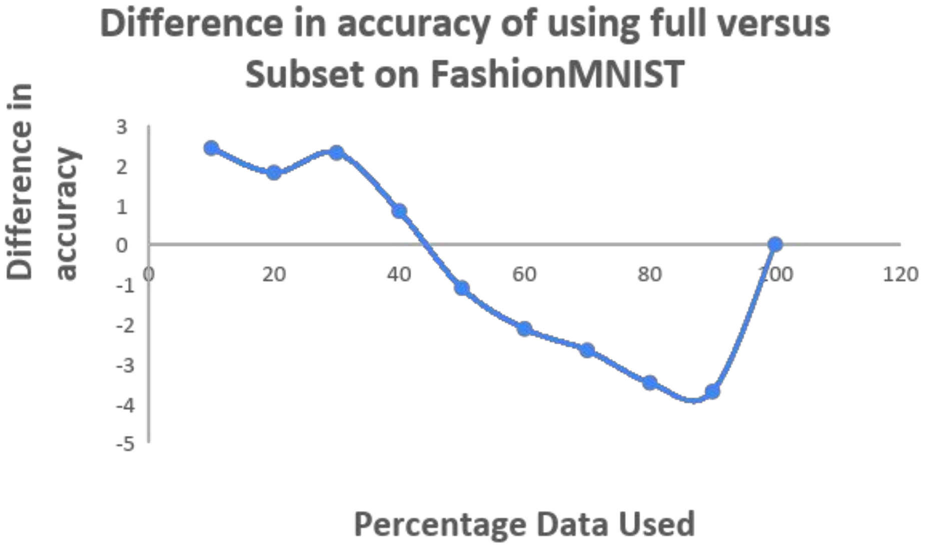


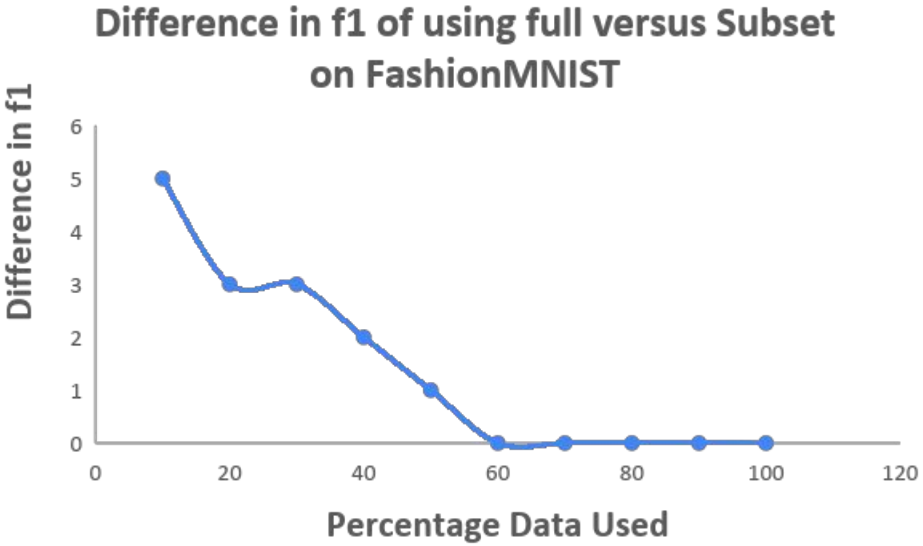
Experiment Results on Fashion MNIST



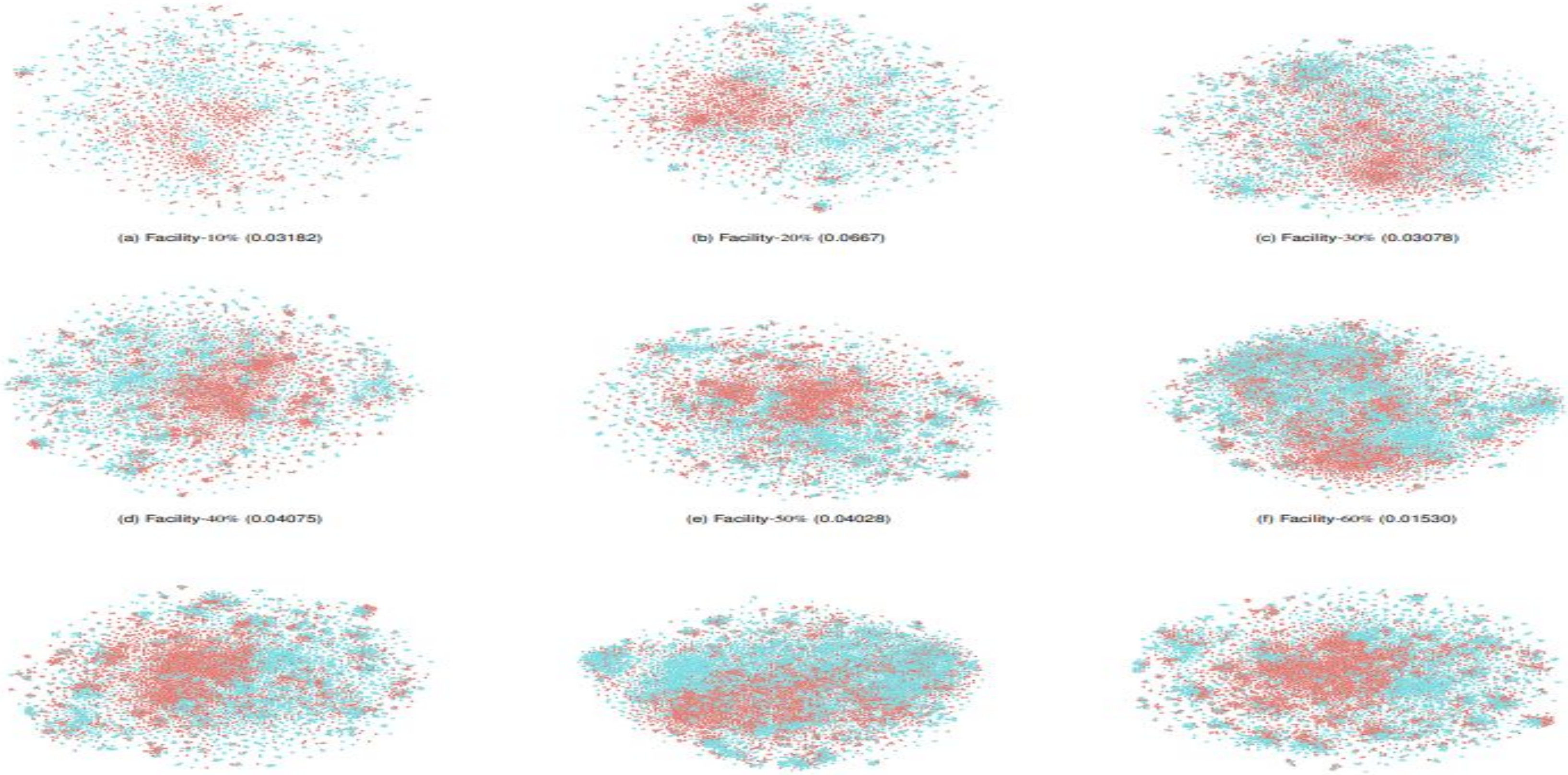


Experiment Results on IMDB Dataset





Multidimensional projections on subsets



(a) Facility-10% (0.03182)

(b) Facility-20% (0.0667)

(c) Facility-30% (0.03078)

(d) Facility-40% (0.04075)

(e) Facility-50% (0.04028)

(f) Facility-60% (0.01530)

(g) Facility-70% (0.05627)

(h) Facilty-80% (0.02321)

(i) Full Dataset-100% (0.04406)

Conclusion

- The choice of the selection matters
- Submodular data selection functions effectively select subsets that maintain acceptable performance on the ground set
- Submodular selections often select better subsets compared to random baseline
- Our two-stage sampling method reduces the memory required to run subset selection algorithms

THANK YOU

END