EE219: Project Report

Clustering

Abdullah-Al-Zubaer Imran

Curtis Crawford

1 Introduction

Clustering alogrithms find groups of data points that have similar representations in a proper space, in unsupervised way. Clustering differs from classification in that without having any prior labelling of the data points. K-means clustering is a clustering technique that interatively groups data points into regions characterized by a set of cluster centroids. Data representation is very crucial for any clustering algorithm like K-means. In this project, we have figured out proper representations of the data points so that we can get efficient and reasonable results from the clustering. Then we performed K-means clutering on the dataset and evaluated performance using different performance measures. Moreover, different preprocess techniques were performed for possible increase in performance of the clusering.

2 Dataset

For this project, we have used "20 Newsgroups" dataset which is a collection of approximately 20,000 documents, partitioned evenly across 20 different newsgroups, each corresponding to a different topic. And each topic is viewed as a class. Since we performed clustering on this dataset, we pretended that the class labels are not available in the dataset.

3 Working Procedures & Results

3.1 Data Representation

In order to find a good representation of the data, the documents were transformed into TF-IDF vectors using min_df = 3. The Tf-IDF matrix dimension: (7882, 27768)

3.2 Clustering

Then we applied K-means clustering with k = 2 to determine the groups or classes the data points belong to, without providing any prior label. For evaluation purpose, we re-labeled data with either 0 for comp-tech or 1 for rec. And compared the clustering results with the known labels.

3.2.1 Performance Measures

In addition to this, we examined several measures to make a concrete comparison of the clustering results. Results from different measures are reported in the following table:

Performance metrics' scores for clustering

Measure	Score
Homogeneity:	0.791324640919
Completeness:	0.79150685559
V-measure:	0.791415737766
Rand:	0.872390054996
Mutual info:	0.79130553663

$$Contigency = \begin{bmatrix} 151 & 3752 \\ 3870 & 109 \end{bmatrix}$$

3.3 Data Preprocessing

As we observe from the clustering result, TF-IDF vector did not yield a good result for K-means clustering. Therefore, we tried with better representations of the data. We performed two dimensionality reduction techniques as the preprocess for K-means clustering.

3.3.1 Dimensionality Reduction

We have used Latent Semantic Indexing (LSI) and Non-negative Matrix Factorization (NMF) for dimensionality reduction. We determined the effective dimension of the data through inspection of the top singular values of the TF-IDF matrix and noticed how many of them are significant in reconstructing the matrix with the truncated SVD representation. We checked what ratio of the variance of the original data is retained after dimensionality reduction. Figure 1 shows the plot of the percent of variance the top r principle components can retain vs. r, for r = 1 to 1000.

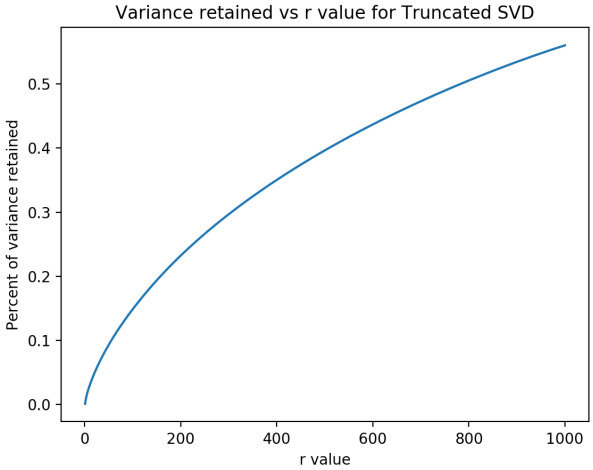


Figure 1: Plot of the Percent of variance retained in PCA vs. r.

For dimensionality reduction, we used LSI and NMF methods. We swept over the parameters for each method (LSI and NMF) to determine the one yielding better results in terms of clustering metrics. All five performance metrics for clustering with different r-values are reported below.

NMF with r = 1

Performance metrics' scores

Measure	Score
Homogeneity:	0.000311084586659
Completeness:	0.00031474279897
V-measure:	0.000312903000955
RAND:	0.000349910576703
Mutual Info:	0.000219562406476

$$Contigency = \begin{bmatrix} 1744 & 2159 \\ 1696 & 2283 \end{bmatrix}$$

SVD with r = 1

Performance metrics' scores

Measure	Score
Homogeneity:	0.000310977615107
Completeness:	0.000314664424341
V-measure:	0.000312810156833
RAND score:	0.000349914958924
Mutual Info:	0.000219455422027

$$Contingency = \begin{bmatrix} 2160 & 1743 \\ 2284 & 1695 \end{bmatrix}$$

NMF with r = 2

Peformance metrics' scores

Measure	Score
Homogeneity:	0.592844515412
Completeness:	0.608067163036
V-measure:	0.600359358773
RAND score:	0.648591716894
Mutual Info:	0.592807239875

$$Contingency = \begin{bmatrix} 731 & 3172 \\ 3943 & 36 \end{bmatrix}$$

SVD with r = 2

Peformance metrics' scores

Measure	Score
Homogeneity:	0.608223241581
Completeness:	0.608333021975
V-measure:	0.608278126825
RAND:	0.713926529273
Mutual Info:	0.608187374307

$$Contingency = \begin{bmatrix} 250 & 3653 \\ 3618 & 361 \end{bmatrix}$$

NMF with r = 3

Performance metrics' scores

Measure	Score
Homogeneity:	0.237561424862
Completeness:	0.317099662339
V-measure:	0.271627663619
RAND score:	0.16950318518
Mutual Info:	0.237491614778

$$Contingency = \begin{bmatrix} 13 & 3890 \\ 1674 & 2305 \end{bmatrix}$$

SVD with r = 3

Performance metrics' scores

Measure	Score
Homogeneity:	0.0353596802034
Completeness:	0.165160546781
V-measure:	0.0582487283625
RAND score:	0.00593193880668
Mutual Info:	0.0352712181601

$$Contingency = \begin{bmatrix} 3635 & 268 \\ 3979 & 0 \end{bmatrix}$$

NMF with r = 5

Performance metrics' scores

Measure	Score
Homogeneity:	0.125884883543
Completeness:	0.127229904183
V-measure:	0.126553820227
RAND score:	0.165339719484
Mutual Info:	0.125804857758

$$Contingency = \begin{bmatrix} 2992 & 911 \\ 1427 & 2552 \end{bmatrix}$$

SVD with r = 5

Performance metrics' scores

Measure	Score
Homogeneity:	0.138545661957
Completeness:	0.154488808534
V-measure:	0.146083525309
RAND score:	0.15259281864
Mutual Info:	0.13846679232

$$Contingency = \begin{bmatrix} 445 & 3458 \\ 2023 & 1956 \end{bmatrix}$$

NMF with r = 10

Performance metrics' scores

Measure	Score
Homogeneity:	0.474595160933
Completeness:	0.513066612395
V-measure:	0.4930816157
RAND score:	0.473136537245
Mutual Info:	0.474547058583

$$Contingency = \begin{bmatrix} 1226 & 2677 \\ 3975 & 4 \end{bmatrix}$$

SVD with r = 10

Performance metrics' scores

Measure	Score
Homogeneity:	0.231788794819
Completeness:	0.319083600677
V-measure:	0.268519547729
RAND score:	0.154588731327
Mutual Info:	0.23171845505

$$Contingency = \begin{bmatrix} 3900 & 3 \\ 2388 & 1591 \end{bmatrix}$$

NMF with r = 20

Performance metrics' scores

Measure	Score
Homogeneity:	0.103775132137
Completeness:	0.213011153692
V-measure:	0.139559454496
RAND score:	0.0388697375327
Mutual Info:	0.103693048241

$$Contingency = \begin{bmatrix} 3894 & 9 \\ 3154 & 825 \end{bmatrix}$$

SVD with r = 20

Performance metrics' scores

Measure	Score
Homogeneity:	0.233028131747
Completeness:	0.320016548166
V-measure:	0.269681134475
RAND score:	0.155989148922
Mutual Info:	0.232957905546

$$Contingency = \begin{bmatrix} 3 & 3900 \\ 1598 & 2381 \end{bmatrix}$$

NMF with r = 50

Peformance metrics' scores

Measure	Score
Homogeneity:	0.0667025153879
Completeness:	0.186835673058
V-measure:	0.0983079466928
RAND score:	0.0152959218258
Mutual Info:	0.0666170072715

$$Contingency = \begin{bmatrix} 3 & 3900 \\ 530 & 3449 \end{bmatrix}$$

SVD with r = 50

Performance metrics' scores

Measure	Score
Homogeneity:	0.774707930719
Completeness:	0.775648956185
V-measure:	0.775178157863
RAND score:	0.856346285004
Mutual Info:	0.774687305158

$$Contingency = \begin{bmatrix} 211 & 3692 \\ 3896 & 83 \end{bmatrix}$$

NMF with r = 100

Peformance metrics' scores

Measure	**Score**
Homogeneity:	2.21983210362e-07
Completeness:	6.94847451879e-06
V-measure:	4.30222097127e-07
RAND score:	-4.45905813813e-07
Mutual Info:	-9.31724050412e-05

$$Contingency = \begin{bmatrix} 13 & 3890 \\ 13 & 3966 \end{bmatrix}$$

SVD with r = 100

Peformance metrics' scores

Measure	Score
Homogeneity:	0.245732969386
Completeness:	0.329585259245
V-measure:	0.281548403613
RAND score:	0.170550013258
Mutual Info:	0.245663907331

$$Contingency = \begin{bmatrix} 3900 & 3 \\ 2310 & 1669 \end{bmatrix}$$

NMF with r = 300

Peformance metrics' scores

Measure	Score
Homogeneity:	0.00256529809666
Completeness:	0.0222595262258
V-measure:	0.00460042089466
RAND score:	-6.92914675877e-05
Mutual Info:	0.00247362159362

$$Contingency = \begin{bmatrix} 3871 & 32 \\ 3889 & 90 \end{bmatrix}$$

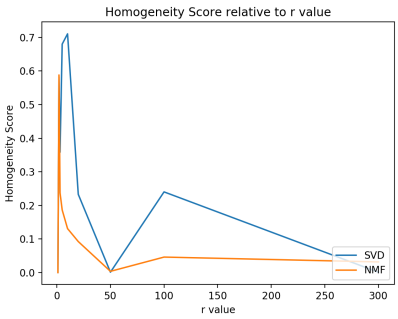
SVD with r = 300

Performance metrics' scores

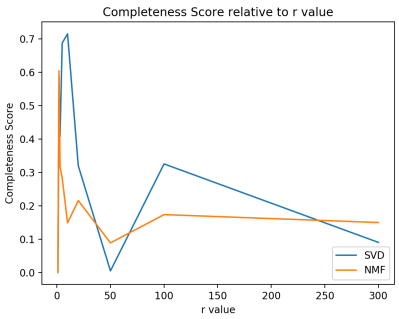
Measure	Score
Homogeneity:	0.241189275662
Completeness:	0.302600706133
V-measure:	0.268427325146
RAND score:	0.197762294913
Mutual Info:	0.24111979987

$$Contingency = \begin{bmatrix} 55 & 3848 \\ 1846 & 2133 \end{bmatrix}$$

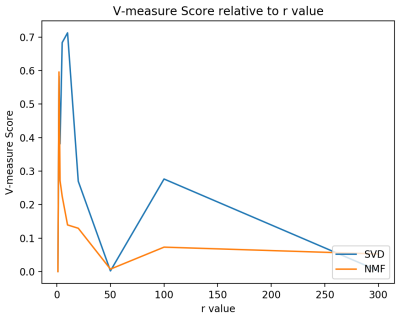
Performance measures for the clustering with different r-values have been visualized in Figure 2. As we can observe non-monotonocity in case of all the measures as r increases. Based on the scores from the measures, the performance of clustering does not seem to be improving or decaying consistently with the increase of r.



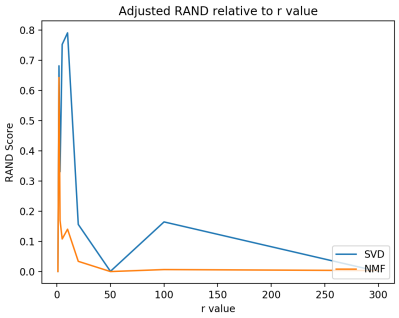
(a) Homogeneity



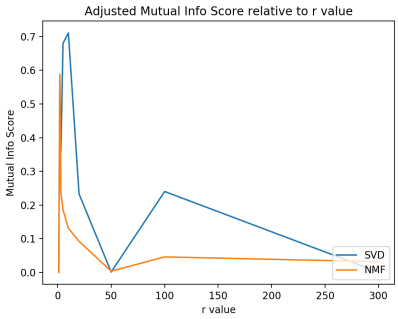
(b) Completeness



(c) V-measure



(d) RAND score

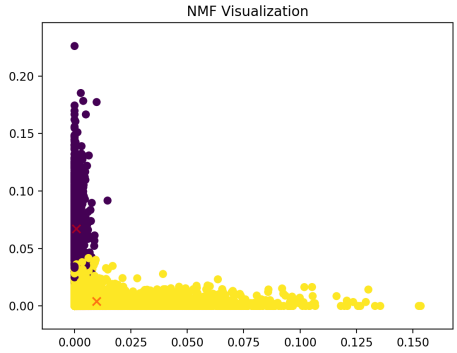


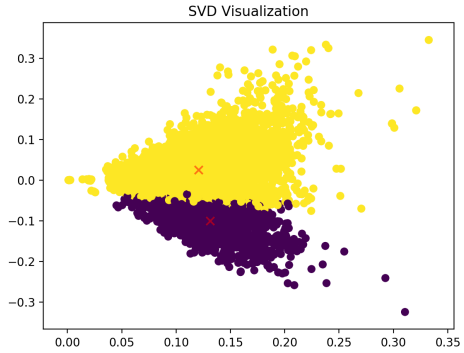
(e) Mutual Info

Figure 2: Performance measures for the clustering using NMF and SVD with different r-values

4 Performance Visualization & Improvement

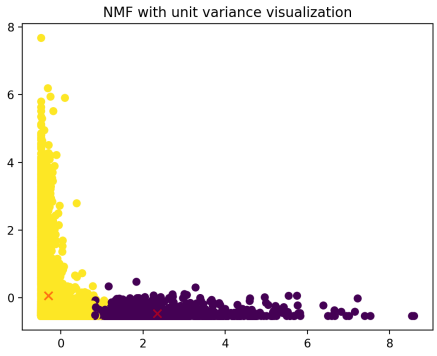
By projecting final data vectors onto 2-dimensional plane and color-coding the classes, the best clustering results from previous part for both SVD and NMF have been visualized in Figure 3. In effort to improve the performance of the clustering, we used three types of transformation techniques: unit variance of all features, logartihmic trnasformation as a non-linear transformation, and the combination of them. The clustering results after these transformations applied have been illustrated in Figure 4 and Figure 5.

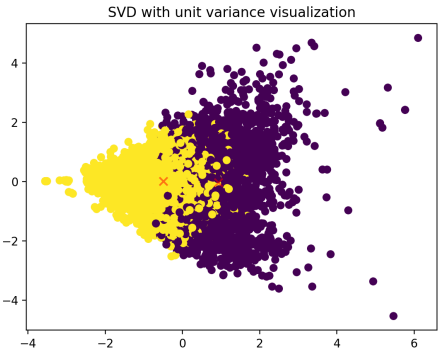




(a) NMF (b) SVD

Figure 3: Best clustering results for NMF and SVD with color-coded classes

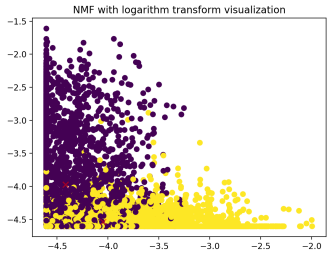




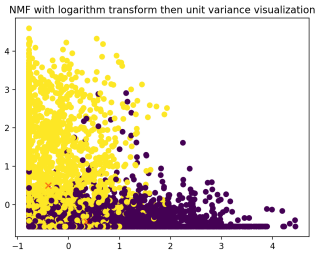
(a) NMF (b) SVD

Figure 4: Clustering results for NMF and SVD with unit variance features

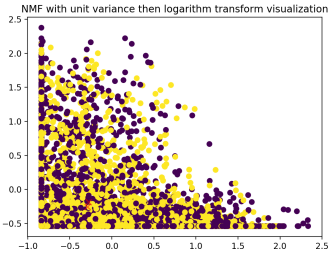
Morevoer, the results from all the performance metrics for the clustering with after transformations have been reported below. All these clearly show the improvement in the performance of the clustering.



(a) NMF with log transformation



(b) NMF with log then scale



(c) NMF with scale then log

Figure 5: Clustering results for NMF with logarithmic, log-scale, and scale-log transformations

SVD with unit variance

Performance metrics' scores

Measure	Score
Homogeneity:	2.01303068525e-05
Completeness:	2.14668865991e-05
V-measure:	2.07771235773e-05
RAND score:	-6.21221803205e-05
Mutual Info:	-7.14202926019e-05

$$Contingency = \begin{bmatrix} 1392 & 2511 \\ 1399 & 2580 \end{bmatrix}$$

NMF with unit variance

Performance metrics' scores

Measure	Score
Homogeneity:	0.558580439281
Completeness:	0.568874068887
V-measure:	0.563680263802
RAND score:	0.635981622287
Mutual Info:	0.558540026933

$$Contingency = \begin{bmatrix} 692 & 3211 \\ 3873 & 106 \end{bmatrix}$$

NMF with non-linear (log) transform

Performance metrics' scores

Measure	Score
Homogeneity:	0.730830451039
Completeness:	0.733210930879
V-measure:	0.732018755671
RAND score:	0.815054275835
Mutual Info:	0.730805808463

$$Contingency = \begin{bmatrix} 3597 & 306 \\ 77 & 3902 \end{bmatrix}$$

NMF with scale then log transform

Performance metrics' scores

Measure	Score
Homogeneity:	0.000902761815731
Completeness:	0.00090693865041
V-measure:	0.00090484541295
RAND score:	0.00117175851602
Mutual Info:	0.000811294034297

$$Contingency = \begin{bmatrix} 1864 & 2039 \\ 1760 & 2219 \end{bmatrix}$$

NMF with log transform then scale

Performance metrics' scores

Measure	Score
Homogeneity:	0.73419145919
Completeness:	0.73616241497
V-measure:	0.735175616083
RAND score:	0.819642953267
Mutual Info:	0.734167124321

$$Contigency = \begin{bmatrix} 289 & 3614 \\ 3895 & 84 \end{bmatrix}$$

Logarithmic transformation may increase the cluserting results because it gives some important insights useful for the clustering. The log transformation works as a filter for PCA. It filters off some dominant trivial effects dominant for PCA, eventually improving the clustering results.

5 Expansion of Dataset into 20 Categories

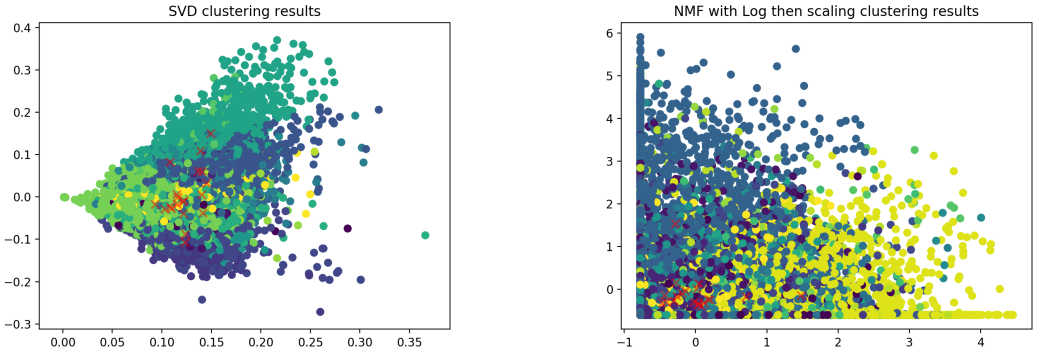
In order to examine how purely we can retrieve all 20 original sub-class labels with clustering, we included all documents and the correspodning terms in the data matrix and figured out proper representation through diemsionality reduction of the TF-IDF representation.

Using the same parameters as in part 1, we tried different dimensions for both truncated SVD and NMF dimensionality reduction. Based on the performance metrics, the best r-values for 20 clusters and 20 sub-classes were found. For 20 clusters, and 20 categories, the best r values have been reported in the table given below.

Best r-values for different performance metrics

Performance metric	**NMF (r-value)**	**SVD (r-value)**
Homegeneity	10	10
Completeness	35	125
V-measure	35	80
RAND	10	10
Mutual info	10	10

Clustering results for both NMF and SVD with logarithmic transformation followed by unit variance have been visualized in Figure 6.



(a) SVD (b) NMF

Figure 6: Clustering results for NMF and SVD with logarithmic transformation followed by unit variance

After trying with different r-values, we used r=35 for NMF, and r=80 for SVD in order to achieve the best clustering performance. Effects of Scaling and Log transform were observed. Therefore,

- SVD
  - Scaling worsened results for r=80
- NMF
  - Scaling worsened results
  - Log improved results
  - Log then scale improved results the most!
  - Scale then log worsened results, but not as bad as just scaling