Feature Reduction for Unsupervised Learning

Meng Wu, Yang Zhao

Abstract

In this project, four unsupervised feature reduction algorithms for clustering problem were investigated and experimented upon two sets of data – handwritten digits data set and the functional magnetic resonance imaging (fMRI) resting state data set. Ratio of sum of squares (RSS), leverage score (LEV), and Laplacian score (LAP) were used to rank the influences of the features in the clustering. Similarity based method were implemented to find largest groups of features that dominate the clustering result. Clustering results were evaluated and compared using both accuracy score and average fisher score.

1 Introduction

An important problem related to machine learning and large dataset mining is selecting a subset effective features from the original data. The goal of feature reduction is to use less number of features while achieving high clustering accuracy or similar clustering results as using all features. Preprocessing the data to obtain a smaller set of representative features, retaining the optimal salient characterstics of the data, not only decreases the processing time but also leads to more compactness of the models learned and better generalization[6]. Feature reduction also helps in improving the prediction performance of the predictors, providing faster and more cost-effective pre-dictors, and providing a better understanding of the underlying process that generated the data. [4]. When class labels of the data are available, supervised feature selection becomes possible and has been studied widely. However, selecting features in unsupervised learning scenario is a much more difficult. It often becomes a NP-hard problem[5].

In this project, we investigated four feature reduction method: (1) ratio of sum of squares (RSS), (2) leverage score (LEV), (3) Laplacian score (LAP) and (4)feature similarity-based method, and experimented with three clustering methods: (a) K-means, (b) EM with mixture of Gaussians, and (c) spectral algorithm on two different datasets.

1.1 Notation

In the report, we represent the n samples data with d-dimensional features as matrix

$$\mathbf{X} = \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \cdots & \\ - & x_n^T & - \end{bmatrix} = [f_1, f_2, ..., f_d] \in \mathbb{R}^{n \times d}, \quad (1)$$

where $x_i \in \mathbb{R}^d$ denotes the ith data point, and $f_j \in \mathbb{R}^n$ denote the jth feature vector. The ith data point belongs to the kth cluster C_k, if $i \in C_k$.

2 Clustering Methods

Numerous clustering methods have been proposed, and different methods have different pros and cons. In the project, we focused on three almost most commonly used clustering method: k-means, expectation maximization with mixture of Gaussian (EM) [1], and spectral algorithm [7, 10]. Different algorithms make different assumption of the data have different clustering results. In the same reason, the feature reduction algorithms may also have different effects on the clustering resulting when the clustering algorithms are not same.

2.1 K-means clustering

Given an unlabeled data set, K-Means is the first clustering method we have chosen. K-means is a popular clustering method because of its simplicity and easiness to implement. In our project, we first tried K-Means with different number of clusters(i.e. k). We used standard

cost function:

$$J(c,\mu) = \sum_{i=1}^{n} \|x_i^{(i\in C_k)} - \mu^k\|^2 \quad (2)$$

Because K-means algorithm may fall into a local optimal, we repeated K-means for 100 times with random initialization, and select the clustering result with the lowest cost function.

2.2 Expectation maximization with mixture of gaussians

Assuming our dataset is a mixture of gaussians, we further used expectation maximization(EM) algorithm to cluster our dataset. EM is another important clustering method in machine learning, which involves iteration in two steps: expectation(E) step and maximization of expected log-likelihood(M) step. We assumed the hidden variable z_i , which follows a multinomial distribution, indicate which of the k Gaussians each x_i had come from. The log likelihood defined as following:

$$l(c,\phi,\mu,\Sigma) = \sum_{i=1}^{n} \log p(x_i^{(i\in C_k)};\mu^k,\Sigma^k) + \log \phi^k \quad (3)$$

where is $p(x_i^{(i\in C_k)};\mu^k,\Sigma^k)$ is the probability of a Gaussian vector. Similar to the above K-Means method, we repeated it 100 times with random initialization.

2.3 Spectral algorithm

Spectral algorithms [7, 13] uses information contained in the eigenvectors of a data affinity (i.e., item-item similarity) matrix to detect structure. Such an approach has proven effective on many tasks, including information retrieval, web search, image segmentation, word class detection and data clustering. The construct a similarity matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$, we used the Gaussian kernel to compute the similarity $S_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{2t^2})$ between the ith and jth data.

To cluster the frame in to k groups, the spectral algorithm is following.

- Compute $\mathbf{D} = \text{diag}(\mathbf{S1})$,
- Compute the Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$
- Compute the first k eigenvectors of $\mathbf{L}$, $u_1,\ldots,u_k$
- Normalized the each row of $[u_1,\ldots,u_k]$
- Cluster using k-means

There are several interesting properties in the spectral algorithm that may be useful to us. The Laplacian matrix $\mathbf{L}$ can be used to compute score for each features resulting a efficient feature selection; the eigenvalues $\mathbf{L}$ indicate how many block like structure are inside the data.

3 Feature Reduction Methods

3.1 Ratio of sums of squares (RSS)

We assume the clustered data followed the model as

$$x_i^{(i\in C_k)} = \mu^k + \varepsilon_i \quad (4)$$

where $\mu^k \in \mathbb{R}^d$ denotes the mean of the kth cluster, and the $\varepsilon_i \sim N(0,\sigma^2\mathbf{I}_d)$ is the random effect of the data. Then the sum-of-squares (SS) for the jth feature have the relation as

$$\begin{aligned} SS_{j,\text{Total}} &= SS_{j,\text{Between}} + SS_{j,\text{Within}} \\ &= \sum_{k=1}^{K} |C_k|(\mu_j - \mu_j^k)^2 + \sum_{i=1}^{n}(x_{ji}^{(i\in C_k)} - \mu_j^k)^2 \end{aligned} \quad (5)$$

Then ratio-of-sum-of-squares (RSS) is defined as

$$RSS_j = \frac{\frac{1}{K-1}SS_{j,\text{Between}}}{\frac{1}{n-K}SS_{j,\text{Within}}}. \quad (6)$$

Large RSS means the feature has better correlation with clustering results. Therefore, we could used the clustering results from all the features to compute the RSS scores, then select the features with the largest RSS values.

Note that, if we assume the μ_j^k is another random variable with model $N(0,\sigma_j^2)$, the the RSS value has an F-statistic ($F_{K-1,n-K}$) for testing $H_0 : \sigma_j^2 = 0$.[11] The p-value of F test for rejecting the H_0 implies the significance of the jth feature in the clustering process. The RSS methods are applied the clustering results from the three clustering algorithms discussed in the previous section.

3.2 Leverage score (LEV)

Boutsidis et. al. presented a novel feature selection algorithm for the k-means clustering problem. [2] Their algorithm is randomized and uses the (normalized) leverage scores to assign probability to the features. The jth leverage score equals the square of the Euclidian norm of the jth row of $\tilde{\mathbf{V}}$.

$$\varphi_j = \|(V_{j1}, V_{j2}, ..., V_{jk})\|^2/K, \quad (7)$$

where $\tilde{\mathbf{V}} = [v_1, v_2, ..., v_K]$ is the matrix that contains the first K right singular vectors of X. The jth leverage

score characterizes the importance of the jth feature with respect to the k-means objective. In the original paper, these scores form a probability distribution over the columns of $\mathbf{X}$. [2] In this project, we simplified it by selecting the features with the largest the LEV scores.

3.3 Laplacian score (LAP)

Laplacian Score (LAP) is fundamentally based on Laplacian eigenmaps and locality preserving projection. The basic idea of LAP is to evaluate the features according to their locality preserving power. [5, 10, 13] The Laplacian Score of the $j$the feature is defined as: [5]

$$\psi_j = \frac{\tilde{f}_j^T \mathbf{D} \tilde{f}_j}{\tilde{f}_j^T \mathbf{L} \tilde{f}_j}, \tag{8}$$

where the vector $\tilde{f}_j$ is

$$\tilde{f}_j = f_j - \frac{f_j^T \mathbf{D1}}{\mathbf{1}^T \mathbf{D1}}, \tag{9}$$

and $\mathbf{L}$ and $\mathbf{D}$ are defined in the spectral algorithm. For a good feature is to feature, the Laplacian Score ψ_j tends to be big. Therefore, we simply selected the features with the smallest ψ_j values.

3.4 Feature similarity (FS)

Several unsupervised feature selection algorithms based on measuring similarity between features have been proposed. [6, 8] Those algorithm aim to remove the redundant features by clustering the features using some similarity measurements. We used two common similarity measurements: mutual information and correlation coefficients to construct feature similarity matrices $\mathbf{A}^{\text{MI}} \in \mathbb{R}^{d \times d}$ and $\mathbf{A}^{\text{CORR}} \in \mathbb{R}^{d \times d}$.

$$\begin{aligned} \mathbf{A}_{ij}^{\text{MI}} &= MI(f_i \| f_j) \\ \mathbf{A}_{ij}^{\text{CORR}} &= \frac{\text{cov}(f_i, f_j)}{\sqrt{\text{var}(f_i)\text{var}(f_j)}}. \end{aligned} \tag{10}$$

We used the feature similarity matrices to find several feature communities (groups) using a modified spectral algorithm. [3] There are two way to select the features based on the founded groups. The first one is to used select the representative feature from each feature group; the second one is select the feature groups with the largest sizes. We chose the second method, because we want to let the clustering result with reduced features close to the results using all features. Thus, the feature groups with the large size are likely to dominate the clustering. The advantages and disadvantages between two approaches need further studies which are beyond the scope of this project.

4 Data Description

4.1 Handwritten Digit Data

In this data set, 1593 handwritten digits from around 80 persons were scanned, stretched in a rectangular box 16x16 in a gray scale of 256 values.Then each pixel of each image was scaled into a bolean (1/0) value using a fixed threshold. Each person wrote on a paper all the digits from 0 to 9, twice. The commitment was to write the digit the first time in the normal way (trying to write each digit accurately) and the second time in a fast way (with no accuracy). (availabel at http://archive.ics.uci.edu/ml/datasets/ Semeion+Handwritten+Digit)

4.2 fMRI Data

In the last few years, many machine learning algorithms have been studied to investigate the brain's functional activities and connectivities on the function magnetic resonance imaging (fMRI) data. [9] Recent studies showed that the spatial patterns of the temporal blood oxygenation level-dependent (BOLD) signals even without explicit stimulation correlations have strong resemblance with many established brain networks, which are generally referred as resting-state network (RSN). The nonstationary correlational structure of RSN provides an opportunity to extract useful information about human brain. However, there are little effort in investing the potential network changes with the resting-state temporal variability.

Liu and Duyn [12] recently reported a method using point process analysis (PPA) to extracting the RSN patterns from relatively brief (a few seconds long) periods of coactivation or codeactivation of regions. They extracted a few fMRI time frames from the representative participant at time points where fMRI signal in the posterior cingulate cortex (PCC) was high. Notably, these single fMRI frames already show coarse resemblance to the DMN pattern. They also extend their method to select the frames based on the two signals: PCC and medial prefrontal cortex (mPFC).

For this project we used fMRI data at resting state from 34 subjects. The total amount of fMRI frames is 8160, each frame are further decompose into 90 features signals. As suggested in Liu's paper, we obtained 122 sample frames when the PCC signal is the 15% largest among every subjects. Therefore, the data can be presented in a 90 by 1244 matrix, with each column representing a sample frame and each row representing a feature. To avoid the bias caused by different amplitude/unit of different features, we first normalized each feature to zero mean and one standard deviation. This may help us

get each feature evenly-weighted, and the energy of each feature could contribute evenly in our clustering process.

5 Results and Discussions

The evaluation of our clustering result is based on two metrics: accuracy for labeled handwritten digit data, and average fisher score for both two data sets. Accuracy is the ratio of correct clustering to the total data points. The averages fisher score is defined as

$$AFS = \frac{1}{d}\sum_{j=1}^{d}\frac{SS_{j,\text{Between}}}{SS_{j,\text{Within}}} \tag{11}$$

5.1 Handwritten Digit Data

Based on the above description, we first plot thefeature scores of the handwritten digital data as a heat map (Figure 1). Different colors in the heat map represents the significance of different feature in terms of the corresponding feature score (i.e. blue: lowest significance; red: highest significance). From the figure we can see, different feature scores vary differently, but most of them indicate the distinguishable difference between different features (i.e. different colors). This figure intuitively shows us the possibility to adopt meaningful feature reduction in our data set.

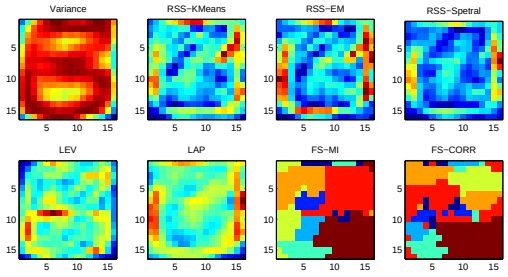


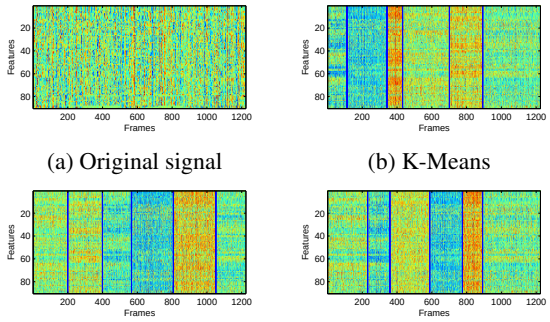
Figure 1: Feature scores for Handwritten digit data.

Then, we experimented with the three different clustering methods by gradually reducing features, based on the different feature scores (Figure 4). Reduction rate is defined as the number of features removed to the number of total features. From the figure, we can see when reduction rate is below 40%, the accuracy remains quite stable around 0.6, which is reasonable for unsupervised clustering methods. Similarly, we also use the average fisher score to evaluate the accuracy, the result is similar. Moreover, we notice that the FS-based methods remain a more stable accuracy and average fisher score in K-means and EM method, which may be a more preferable combination in the real-world application.

5.2 fMRI Data

We chose $k = 6$ for clustering the fMRI data using the three different algorithms as described above. The clusters in all three algorithms have very obvious pattern within each cluster. The clustering agreements between three algorithm are about 60 % (Figure 2). The top 50% and top 75% significant features (brain region) in 10 slices are also shown in Figure 3.

Using similar analysis approach, we show the relation between the feature reduction and average fisher score of fMRI data in Figure 5. Again, the combination of FS-based methods and K-means/EM algorithm achieves a more reasonable average fisher score, demonstrating a more preferable approach in real world application.



(a) Original signal (b) K-Means

(c) EM (d) Spectral

Figure 2: Clustering results of the fMRI data.

6 Conclusion

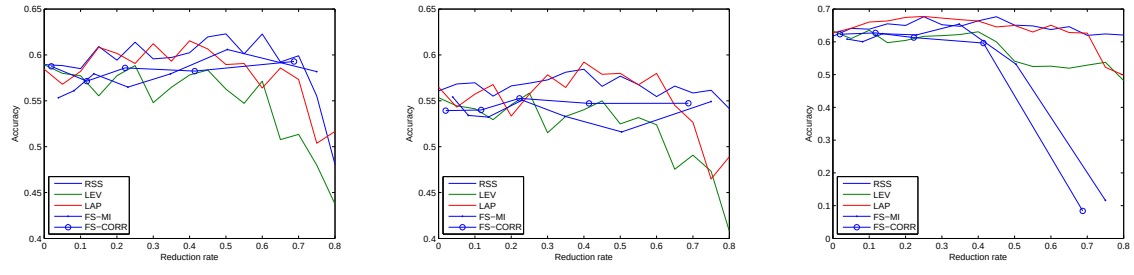
In this project, we focused on a particular unsupervised feature reduction problem, and have investigated four unsupervised feature reduction algorithms: RSS, LEV, LAP, and FS on three unsupervised learning methods: K-means, EM, spectral algorithms. Experiment on the labeled human handwritten digit data has proved the effectiveness of our feature reduction algorithms, and the application on the fMRI data presents a promising application in real world.

7 Acknowledgments

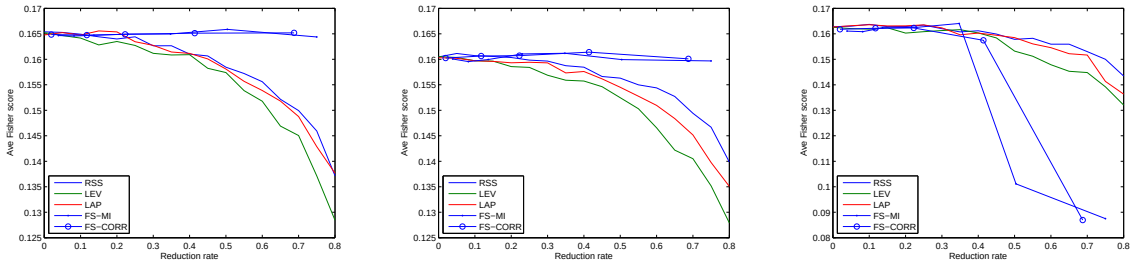
The fMRI data used in this project is provided by Jingyuan Chen from the Radiological Science Lab, Stanford School of Medicine.

References

[1] BILMES, J. A. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute 4*, 510 (1998), 126.

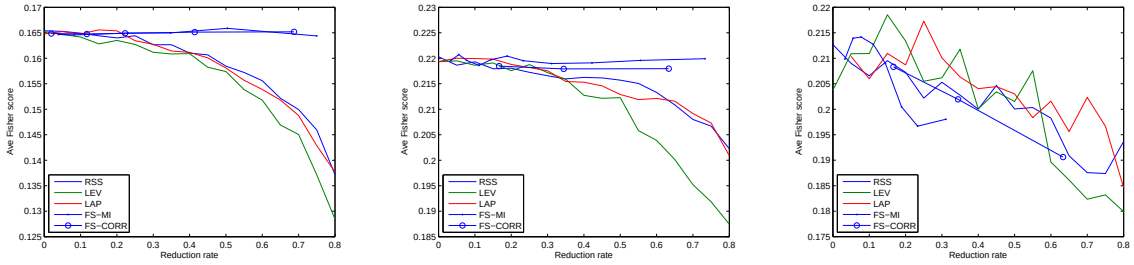


(a) K-means (b) EM (c) Spectral



(d) K-means (e) EM (f) Spectral

Figure 4: Clustering accuracy and average Fisher score with feature reduction using handwriting digits data set.



(a) K-means (b) EM (c) Spectral

Figure 5: Average Fisher score with feature reduction using fMRI data set.

[2] BOUTSIDIS, C., DRINEAS, P., AND MAHONEY, M. W. Unsupervised feature selection for the k-means clustering problem. In *Advances in Neural Information Processing Systems* (2009), pp. 153–161.

[3] FORTUNATO, S. Community detection in graphs. *Physics Reports 486*, 3 (2010), 75–174.

[4] GUYON, I., AND ELISSEEFF, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research 3* (2003), 1157–1182.

[5] HE, X., CAI, D., AND NIYOGI, P. Laplacian score for feature selection. In *Advances in neural information processing systems* (2005), pp. 507–514.

[6] MITRA, P., MURTHY, C., AND PAL, S. K. Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence 24*, 3 (2002), 301–312.

[7] NG, A. Y., JORDAN, M. I., AND WEISS, Y. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS* (2001), MIT Press, pp. 849–856.

[8] PENG, H., LONG, F., AND DING, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*, 8 (2005), 1226–1238.

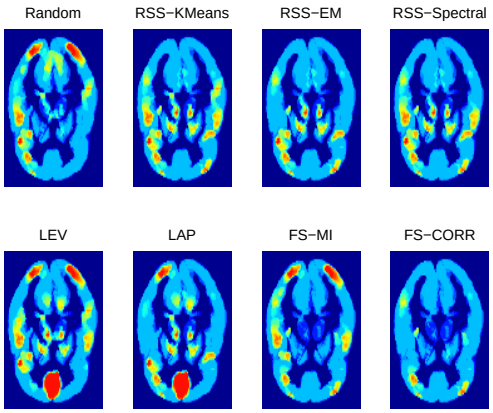
[9] PEREIRA, F., MITCHELL, T., AND BOTVINICK, M. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage 45*, 1 (2009), S199–S209.

[10] VON LUXBURG, U. A tutorial on spectral clustering. *Statistics and computing 17*, 4 (2007), 395–416.

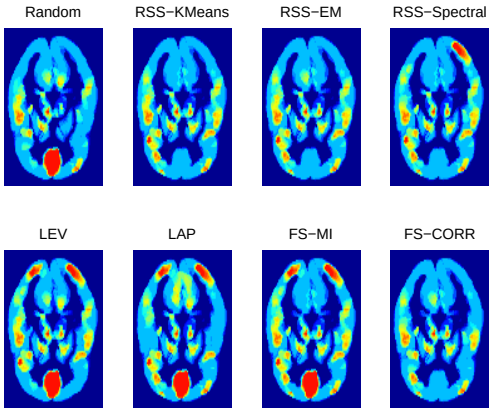
[11] WEISBERG, S. *Applied linear regression*, vol. 528. Wiley. com, 2005.

[12] XIAO LIU, J. H. D. Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences of the United States of America* (2013).

[13] ZHAO, Z., AND LIU, H. Spectral feature selection for supervised and unsupervised learning. In *In ICML* (2007).



(a) Top 50 % regions



(b) Top 75 % regions

Figure 3: Most significant brain regions using differenct selection methods.