Unsupervised Single and Multiple View Feature Extraction for High Dimensional Data Clustering

Jayashree

Under the Guidance of

Dr.Shivaprakash T.
Professor,
Department of Computer Science and Engineering
Vijaya Vittala Institute of Technology,Bangalore

February 23, 2021

Outline

- Introduction
- Literature survey
- Objective
- Methodology
- System Framework
- Work carried out so far
  - Feature extraction algorithms-results
  - Optimal value for k-results
- Further Work to be carried out
- Conclusion

Introdution

- High dimensional large volume of data is challenging for processing
- Dimensionality reduction reduces the challenge
    - Feature extraction method
    - Feature selection method
- Feature extraction- two categories
    - Supervised learning
    - Unsupervised learning
- Typical unsupervised feature extraction method depends on
    - Graph construction method and its fixed graph without learning mechanism
    - Lack of structure information
- To overcome dependency
    - Feature Extraction Structured Graph(FESG)[1] is effective feature extraction method
    - Proposed method- automatically identify the number of clusters
    - Analysis of convergence behavior of the Multiview Feature Extraction Structured Graph(MEFSG) algorithm

Literature survey

Year & Author	Source of Reference	Title	Methods	Performance	Drawback
2016 Zhuge et al.	IEEE	Unsupervised Feature Extraction using a Learned Graph with Clustering Structure	LGCS	-Learn both transformation matrix and ideal graph -Effective projection ability and structured graph	Parameter determination
2017 Zhuge et al.	IEEE	Unsupervised Single and Multiple Views Feature Extraction with Structured Graph	FESG MFESG	Framework for feature extraction	Number of cluster determination
2018 Shi et al.	ELSEVIER	Unsupervised multi-view feature extraction with dynamic graph learning	UMFE-DGL	-Dynamic graph construction -Deep feature co-relation of different view	Performance depends on range of parameters
2018 Yin et al.	Springer	Multi-view clustering via spectral embedding fusion	MVSEF	Objective function to find fusional embedding of global and local structure informtion	complexity of clustering
2019 Shi et al.	ELSEVIER	Auto-weighted multi-view clustering via spectral embedding	AMCSE	Avoids 2 step methods of clustering but it learns clustering structure and obtain the clustering results	It can not deal large scale dataset
2020 Fang et al.	IEEE	ANIMC: A Soft Framework for Auto-weighted Noisy and Incomplete Multi-view Clustering	ANIMC	It automatically learns a proper weight for each view, so that reducing the influence of noises.	Limitation on incomplete multi-view clustering on large-scale data with noises.

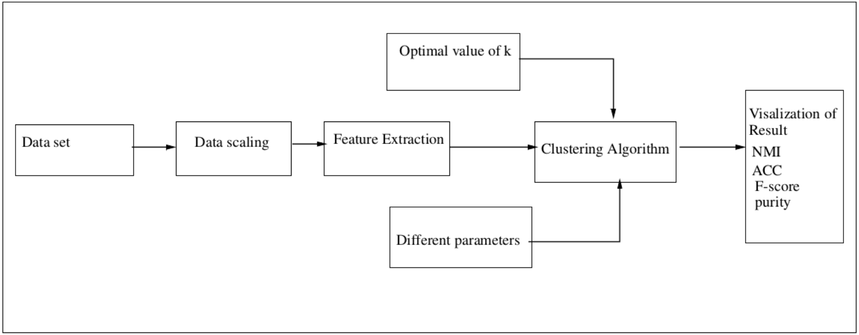
Objective

- The proposed frame work uses FESG and MFESG algorithms
  - graph construction
  - learn graph using dynamic technique
- Automatically identify the number of clusters
- Analysis of convergence behavior of the general algorithm MEFSG
- Apply this strategy to other multi view methods

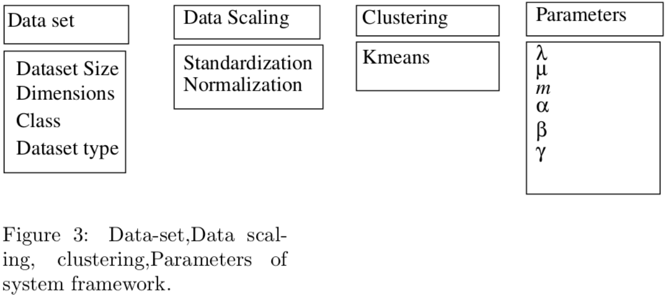
Methodology

- FESG method, adopts
  - The initial graph construction
  - Parameters of number of clusters has to set in the framework.
  - Use k-means to cluster the embedding data.
  - Repeat experiment with different data for performance results.
- There are six different methods,
  - Show convergence behavior
  - Get the clustering results of K-means on different data with different numbers of extracted features
  - Clustering results of K-means on multi-view datasets to test the projection ability of MFESG.
  - Show some clustering results of ideal structured graph matrix learned by FESG and MFESG
  - Results with different parameters.
  - Some results of other methods within the frameworks

System Framework



Dataset,scalling and clustering



Data set

Dataset Size
Dimensions
Class
Dataset type

Data Scaling

Standardization
Normalization

Clustering

Kmeans

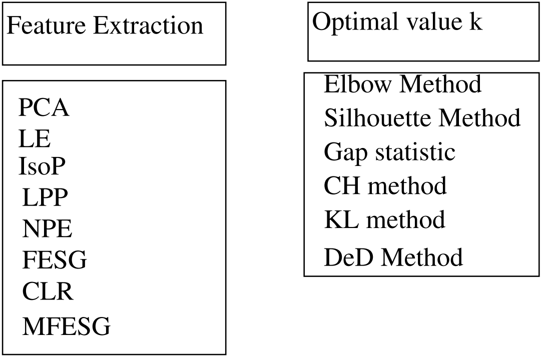
Parameters

λ
μ
m
α
β
γ

Figure 3: Data-set,Data scaling, clustering,Parameters of system framework.

Figure 1: Dataset,scalling and clustering paremeters

Existing algorithms



Feature Extraction

PCA
LE
IsoP
LPP
NPE
FESG
CLR
MFESG

Optimal value k

Elbow Method
Silhouette Method
Gap statistic
CH method
KL method
DeD Method

Figure 2: Different algorithms for Feature extraction and Optimal k value

Work carried out so far

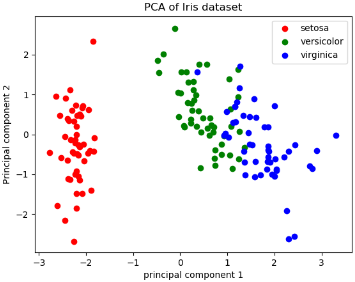
- Literature survey on various algorithms of dimensionality reduction including feature selection and feature extraction.
- Submitted survey paper and accepted in DeepLUDA 2020 conference in "Hyatt Regency Tianjin East Tianjin, China"
- Literature survey on different methods to find the optimal value for k(number of clusters) for the clustering algorithm.
- Learned latex, xfig diagram tools.

- Created instance in Alibaba cloud with OS Ubuntu 18.04 64-bit and 1GB memory.
- Installed Anaconda along with scikit packages.
- Executed few algorithms such as PCA,ICA, ISoP etc.
- Executed some of the existing algorithms for finding optimal number of clusters

Feature extraction algorithms-results

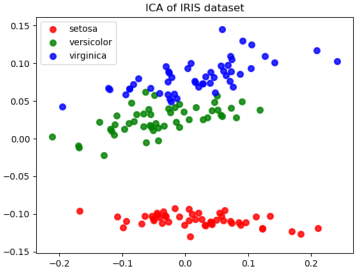
- The Feature extraction algorithms of iris data-set which is build in with sklearn package.

Feature extraction algorithms-PCA



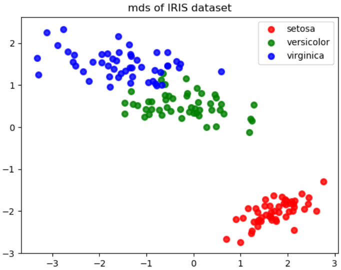
- Uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables.

Independent component analysis(ICA)



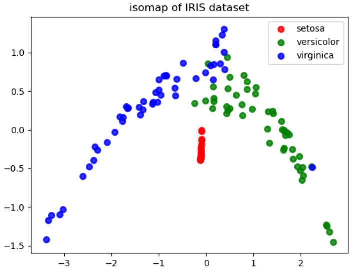
- ICA is an extension of the PCA
- ICA is based on the assumption that source signals are statistically independent

Multidimensional scaling (MDS)



- Works when the data is embedded linearly, or nearly linearly, within the observation space
- MDS algorithms employing Euclidean principles

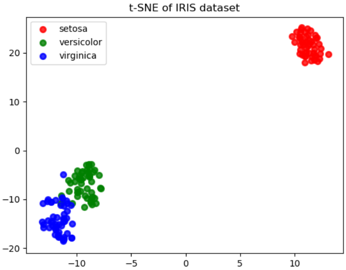
Isomap



There are three steps for Isomap:

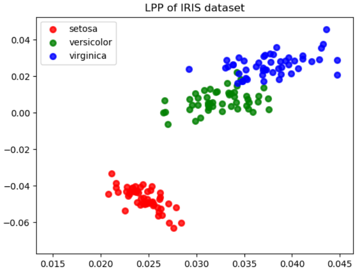
1. Construct neighborhood graph on the manifold.
2. Compute the shortest path between pairwise points by geodesic distances.
3. Construct low-dimensional embedding by applying MDS.

t-Distributed Stochastic Neighbor Embedding(t-SNE)



- A non-linear dimensionality reduction algorithm
- reduces the data dimensions into two or more dimensions from hundreds or thousands of original data-set dimensions.

Locality preserving projection (LPP)

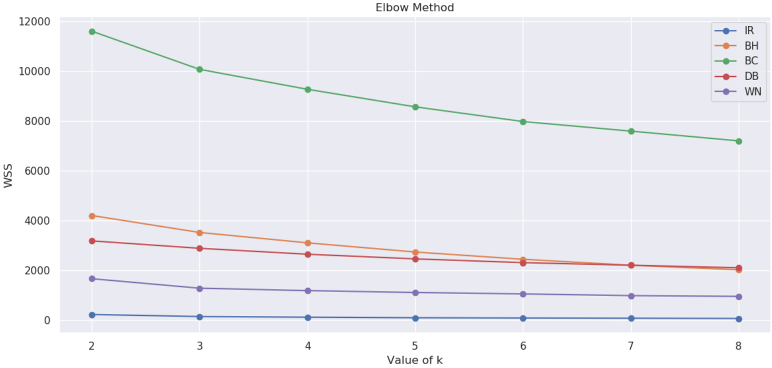


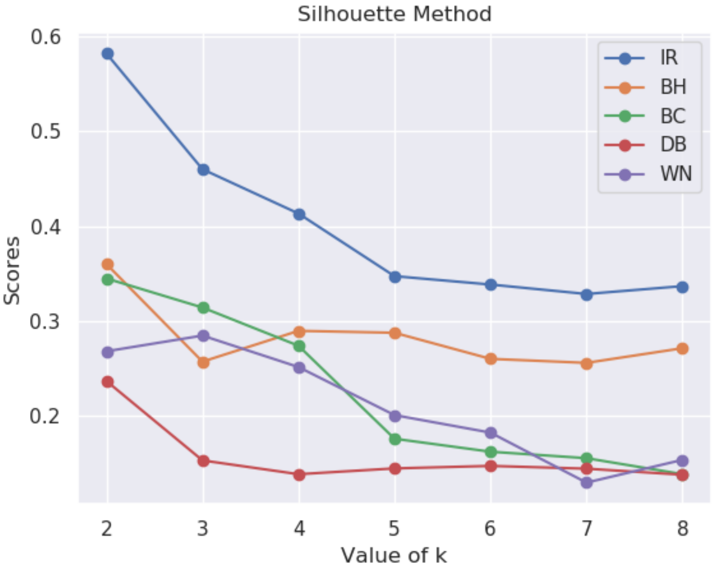
- Constructs a graph incorporating neighborhood information of the data set
- By using Laplacian of the graph, calculate a transformation matrix which maps the data points to a subspace data.

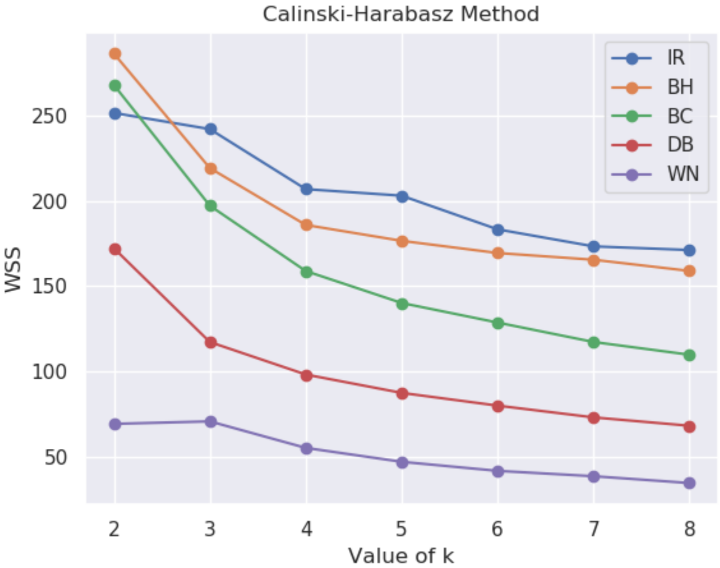
Optimal value for k-results

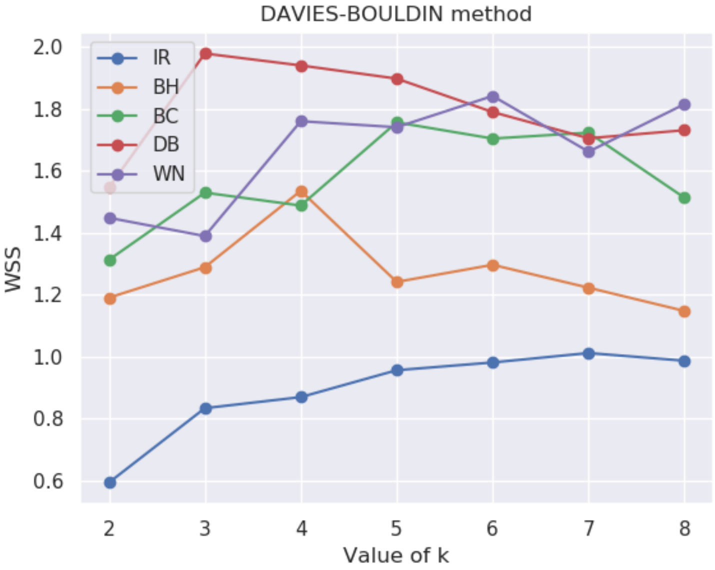
Table 1: Characteristics of the Data set

Dataset	Instances	Attributes	Clusters
Boston house prices dataset(BH)	506	14	
Iris plants dataset(IR)	150	4	3
Diabetes dataset(DB)	442	10	2
Wine recognition dataset(WN)	178	13	3
Breast cancer(diagnostic) dataset(BR)	169	30	2


Elbow Method
12000
10000
8000
6000
4000
2000
0
WSS
2
3
4
5
6
7
8
Value of k
IR
BH
BC
DB
WN


Silhouette Method
0.6
0.5
0.4
0.3
0.2
Scores
2
3
4
5
6
7
8
Value of k
IR
BH
BC
DB
WN


Calinski-Harabasz Method
IR
BH
BC
DB
WN
WSS
250
200
150
100
50
2
3
4
5
6
7
8
Value of k


DAVIES-BOULDIN method
IR
BH
BC
DB
WN
WSS
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
2
3
4
5
6
7
8
Value of k

Further Work to be carried out

1. Find the suitable algorithm of finding k to embed with different feature extraction algorithm.
2. Analyze the different feature extraction algorithms concerning the time of execution along with different data-sets.
3. Construct Multi view features using different methods and analyzes convergence behavior of MFESG
4. Construct the framework with a clustering algorithm.
5. Analyze the clustering algorithm with different data sets by F-score, NMI(Normalized Mutual Information) and mean ACC(Clustering Accuracy).

Conclusion

- Proposed Unsupervised feature extraction technique uses
  - learned graph construction method
  - structured graph/dynamic graph
  - clustering technique produce effective results are expected

  - Automatically set some variables
  - Analyze the convergence behavior of algorithm
- Application of FESG and MFESG to other multi view methods

Bibliography

Wenzhang Zhuge, Feiping Nie, Chenping Hou, and Dongyun Yi.
Unsupervised single and multiple views feature extraction with structured graph.
IEEE Transactions on Knowledge and Data Engineering, 29(10):2347–2359, 2017.

Xiang Fang, Yuchong Hu, Pan Zhou, Xiao-Yang Liu, and Dapeng Oliver Wu.
Animc: A soft framework for auto-weighted noisy and incomplete multi-view clustering.
arXiv preprint arXiv:2011.10331, 2020.

Shaojun Shi, Feiping Nie, Rong Wang, and Xuelong Li.
Auto-weighted multi-view clustering via spectral embedding.
Neurocomputing, 399:369–379, 2020.

Hongwei Yin, Fanzhang Li, Li Zhang, and Zhao Zhang.
Multi-view clustering via spectral embedding fusion.
Soft Computing, 23(1):343–356, 2019.

Thank you