Project Work

Sean Deloddere - 914691
Giulio Marcon - 914756
CS-E4650 - Methods for Data Mining

December 7, 2020

1 Introduction

In this report we will discuss our approaches and results for clustering 2 data sets, `Gene data` and `MS data`. We evaluated the goodness of our clustering methods with normalized mutual information (NMI), between the labels provided with the data and our inferred labels.

2 Methods

To cluster our data we used K-Means, Hierarchical and Spectral clustering. We opted for these methods as we had experience with them from the home assignments, and found out that they were able to achieve high results when tuned correctly.

2.1 Pre-processing

When testing our models, we used several versions of the data:

- No pre-processing
- **Standardization**: We used the `StandardScaler()` method from `Scikit-learn`, a Python library, to standardize the data using the mean and variance: $z = (x - \mu)/\sigma$, with x being the original data point, μ being the mean and σ the standard deviation [1].
- **Maximum scaling**: We used the `MaxAbsScaler()` method from `Scikit-learn` to scale each feature by its maximum absolute value [1].

2.2 K-Means

`K-Means clustering` is a clustering method that aims to cluster data by iteratively assigning points to the cluster such that the sum of the squared distance between the point and the cluster's centroid is minimized. For our implementation of K-Means we employed `Scikit-learn` [1]. We calculated the NMI of labels predicted by K-Means for a number of clusters ranging from 2 to 9, and for 3 differently processed inputs, see Section. 2.1.

2.3 Hierarchical

`Agglomerative clustering`, the type of hierarchical clustering we implemented, is a bottom-up approach to clustering, where each data point starts in its own clusters, and pairs of clusters are merged greedily based on minimally increasing a given linkage distance. For our implementation we used Scikit-learn [1]. We tested clustering for 4 types of linkage distances:

- **Ward**: minimizes the variance of the clusters being merged.
- **Average**: uses the average of the distances of each data point of the 2 clusters.
- **Complete**: uses the maximum distances between all data points of the 2 clusters.
- **Single**: uses the minimum distances between al data points of the 2 clusters.

For all 4 of these linkage distances, we calculated the NMI of labels predicted for a number of clusters ranging from 2 to 9, and for 3 differently processed inputs.

2.4 Spectral

`Spectral clustering` is a clustering methods that builds clusters using the eigenvalues of a similarity, also called affinity, matrix to perform dimensionality reduction before clustering in fewer dimensions. We used `Scikit-learn` for our implementation [1]. We used 4 types of similarity matrices:

- **Nearest Neighbors**: Type already included in the `Scikit-learn` `SpectralCluster` method. Constructs the matrix by computing a graph of the nearest neighbors. Takes the amount of nearest neighbors as an input, through experimental methods we determined to let this value range from 5 to 9 while testing.
- **RBF**: Also a default type in the `Scikit-learn` `SpectralCluster` method. Constructs the matrix using a radial basis function (RBF) kernel.
- **Cosine Similarity**: The cosine of the angle between two (non-zero) vectors of an inner product space. We calculated this using `Scikit-learn` [1].
- **Euclidean Similarity**: Similarity measure based on the Euclidean distance between 2 points. It was calculated by applying the following formula: $1/(1+d)$ with d being the Euclidean distance between 2 data points. The Euclidean distance was calculated using `Scikit-learn` [1].

We calculated the NMI of labels predicted for a number of clusters ranging from 2 to 8, for all 4 of these similarity measures, and for 3 differently processed inputs.

3 Results

3.1 Gene data

3.1.1 K-Means

The results for `K-Means` applied to the `Gene data` can be seen in Table. 1. The performance of the clustering is the highest for 6 clusters and no pre-processing, with an nmi of 0.883.

3.1.2 Hierarchical

In Table. 2 the results for `Agglomerative clustering` applied to the `Gene data` are displayed. The highest scoring linkage criterion is `Ward` with an NMI of 0.918 for 6 clusters and no pre-processing.

3.1.3 Spectral

Table. 3 contains the results for `Spectral clustering` applied to the `Gene data`. The highest score, 0.985, is achieved by using Nearest Neighbors, no pre-processing, 5 neighbors and 5 clusters.

3.2 MS data

3.2.1 K-Means

The results for `K-Means` applied to the `MS data` can be seen in Table. 4. The performance of the clustering is the highest for 2 clusters and no pre-processing, with an nmi of 0.499.

3.2.2 Hierarchical

In Table. 5 the results for `Agglomerative clustering` applied to the `MS data` are displayed. The highest scoring linkage criterion is `Ward` with an NMI of 0.474 for 5 clusters and no pre-processing.

3.2.3 Spectral

Table. 6 contains the results for `Spectral clustering` applied to the `MS data`. The highest score, 0.915, is achieved by using a Euclidean Similarity matrix, Maximum scaling and 5 clusters.

4 Conclusions

The highest performing model for `Gene data` was `Spectral clustering` with Nearest Neighbors, no pre-processing, 5 neighbors and 5 clusters, resulting in an `NMI` of 0.985. The highest performing model for `MS data` was `Spectral clustering` with a Euclidean Similarity matrix, Maximum scaling and 5 clusters, resulting in an `NMI` of 0.915.

References

[1] "scikit-learn." `https://scikit-learn.org/`, 2020. Accessed: 7.12.2020.

5 Appendix

K-Means		**Gene data**
Pre-processing	# of clusters	nmi
No pre-processing	6	0.883
	5	0.857
	7	0.856
StandardScaler	5	0.798
	8	0.782
	7	0.774
MaxAbsScaler	5	0.846
	7	0.844
	6	0.826

Table 1: K-means configurations for each form of applied pre-processing with 3 highest nmi scores for Gene data.

Hierarchical						**Gene data**
Linkage	Ward			Complete		
	Pre-processing	# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing	6	0.918	No pre-processing	8	0.610
		7	0.888		7	0.600
		5	0.878		6	0.540
	StandardScaler	7	0.838	StandardScaler	8	0.508
		8	0.828		7	0.494
		6	0.817		6	0.320
	MaxAbsScaler	7	0.888	MaxAbsScaler	8	0.618
		6	0.851		7	0.528
		5	0.848		5	0.302
Linkage	Average			Single		
	Pre-processing	# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing	8	0.046	No pre-processing	8	0.036
		7	0.046		7	0.035
		6	0.040		6	0.030
	StandardScaler	8	0.077	StandardScaler	8	0.036
		7	0.074		7	0.032
		6	0.038		6	0.030
	MaxAbsScaler	8	0.049	MaxAbsScaler	8	0.036
		7	0.044		7	0.032
		6	0.040		6	0.030

Table 2: Hierarchical configurations for linkage method and each form of applied pre-processing with 3 highest nmi scores for Gene data.

Spectral						**Gene data**
Affinity	Nearest Neighbor			RBF		
	Pre-processing	# of neighbors	# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing	5	5	0.985	No pre-processing	7	0.009
		5	7	0.976		6	0.009
		6	5	0.937		5	0.005
	StandardScaler	5	7	0.967	StandardScaler	3	0.009
		5	5	0.949		7	0.008
		6	7	0.925		4	0.007
	MaxAbsScaler	5	7	0.977	MaxAbsScaler	7	0.031
		5	5	0.976		6	0.030
		6	5	0.932		5	0.028
Affinity	Cosine Similarity				Euclidean Similarity		
	Pre-processing		# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing		6	0.778	No pre-processing	5	0.806
			4	0.754		6	0.733
			7	0.753		7	0.726
					StandardScaler	7	0.663
						3	0.651
						6	0.619
	MaxAbsScaler		6	0.754	MaxAbsScaler	5	0.759
			7	0.730		6	0.712
			5	0.713		4	0.701

Table 3: Spectral configurations for different affinities and each form of applied pre-processing with 3 highest nmi scores for Gene data.

K-Means		**MS data**
Pre-processing	# of clusters	nmi
No pre-processing	2	0.499
	4	0.352
	3	0.272
StandardScaler	1	0.250
	8	0.150
	3	0.113
MaxAbsScaler	1	0.250
	4	0.230
	6	0.169

Table 4: K-means configurations for each form of applied pre-processing with 3 highest nmi scores for MS data.

Hierarchical						**MS data**
Linkage	Ward			Complete		
	Pre-processing	# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing	5	0.474	No pre-processing	8	0.268
		6	0.446		1	0.250
		7	0.428		5	0.248
	StandardScaler	6	0.384	StandardScaler	7	0.256
		7	0.370		8	0.251
		8	0.361		1	0.250
	MaxAbsScaler	4	0.583	MaxAbsScaler	1	0.250
		5	0.544		7	0.248
		6	0.505		6	0.245
Linkage	Average			Single		
	Pre-processing	# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing	1	0.250	No pre-processing	1	0.250
		8	0.036		8	0.036
		7	0.033		7	0.033
	StandardScaler	1	0.250	StandardScaler	1	0.250
		8	0.036		8	0.035
		7	0.033		7	0.032
	MaxAbsScaler	1	0.250	MaxAbsScaler	1	0.250
		8	0.036		8	0.036
		7	0.033		7	0.033

Table 5: Hierarchical configurations for linkage method and each form of applied pre-processing with 3 highest nmi scores for MS data.

Spectral							**MS data**
Affinity	Nearest Neighbor				RBF		
	Pre-processing	# of neighbors	# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing	5	4	0.043	No pre-processing	4	0.005
		7	4	0.043		5	0.004
		5	5	0.022		7	0.004
	StandardScaler	5	4	0.043	StandardScaler	3	0.024
		7	4	0.028		2	0.013
		5	5	0.020		4	0.013
	MaxAbsScaler	5	4	0.043	MaxAbsScaler	7	0.089
		7	4	0.032		6	0.073
		5	5	0.021		5	0.064
Affinity	Cosine Similarity				Euclidean Similarity		
	Pre-processing		# of clusters	NMI	Pre-processing	# of clusters	NMI
	No pre-processing		3	0.848	No pre-processing	6	0.880
			2	0.712		7	0.858
			5	0.671		3	0.582
					StandardScaler	6	0.901
						7	0.717
						5	0.578
	MaxAbsScaler		3	0.875	MaxAbsScaler	5	0.915
			2	0.744		6	0.887
			4	0.683		7	0.729

Table 6: Spectral configurations for different affinities and each form of applied pre-processing with 3 highest nmi scores for MS data.