DUSC: Dimensionality Unbiased Subspace Clustering

GROUP - 4

Introduction : Subspace Clustering

- Clustering aims at grouping data such that objects within groups are similar and objects in different groups are dissimilar.
- Relevant attributes may not be uniform globally across all clusters.
- In noisy data or data with many attributes, clusters are often hidden in subspace of the attributes and do not show up across the full attribute space.
- For these applications, subspace clustering aims at detecting clusters in any subspace.

Problem in Subspace clustering

- **Existing Subspace clustering methods don’t consider dimensionality bias.**
- **Dimensionality Bias :** As Dimensions increases, the Avg distances between objects increase, cluster radii grows. The correlation between density of an object and dimensionality of subspace is dimensionality Bias.
- As the dimensionality of subspace varies, approaches which do not take this into account
  - **fail to separate clusters from noise.**
  - **fail to separate dense from sparse regions across subspaces of different dimensionality.**

Our Approach..

Dimensionality Unbiased Subspace Clustering

Density Measure

W : R -> R weighting function , subspace S , density measure is :

$$\varphi^{\mathbf{S}}(o) = \sum_{p \in \mathcal{N}_{\varepsilon}^{\mathbf{S}}(o)} \mathcal{W}\left(\|o - p\|^{\mathbf{S}}\right)$$

A data point o in S is dense , if ϕS(o) ≥ threshold.

Weighting function should monotonically decrease with increase in distance. Here we are using **Epanechnikov kernel.**

Dimensionality Bias still exists in density measure

- As Dimensions increases
  - Avg. distances between objects increase, cluster radii grows
  - Expected density within the area of influence decreases.

- **Problems :**
  - **Difficult to fix the common density threshold for both low and high** dimensional subspaces.
  - Low Threshold : Excess pseudo clusters are formed in low dimensional space.
  - High Threshold : Clusters in the high dimensional space cannot be identified.

Solution - Unbiased density measure

Density measure such that its expected density is same in any two subspaces.

$$\forall\ S1\ ,\ S2 : E\,[\varphi^{\,S\,1}] = E[\varphi^{\,S\,2}]$$

One of the solutions :

Normalise the density with its expected density α(S).

$$E[\frac{1}{E[\varphi^{\mathbf{S}}]}\varphi^{\mathbf{S}}] = \frac{1}{E[\varphi^{\mathbf{S}}]}E[\varphi^{\mathbf{S}}] = 1$$

Then its value is 1 (almost comparable) in all the subspaces.

The normalized density measure is thus given by,

$$\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o) \;\; with$$

$$\alpha(\mathbf{S}) = E_{\mathbf{S}}\left[\varphi^{\mathbf{S}}(o)\right] = \frac{2n\varepsilon^{|\mathbf{S}|}c_{|\mathbf{S}|}}{\mathbf{v}^{|\mathbf{S}|}(|\mathbf{S}|+2)}$$

Problems Handled : DUSC Subspace clustering

- **Intuitive density threshold : F**
  - Instead of absolute thresholds, only factor by which the expected density to be exceeded is used. $\varphi^S(o) \geq F \cdot \alpha(S)$
  - So, same F value can be used across all subspaces.

Problems Handled : DUSC Subspace clustering

- **Redundancy :**
  - Remove redundant clusters that essentially contain the same information repeated in different dimensionalities.
  - A cluster is redundant if most of the objects in the cluster in subspace **S** are also in another cluster in a higher dimensional subspace **S'** $\supset$ **S**.
  - **Not redundant**: $\neg \exists (C, S)$ subspace cluster with $C' \subseteq C \wedge S \subset S' \wedge |C| \geq r \cdot |C'|$
  - Parameter r to specifies the degree of redundancy acceptable to the user.

Cluster Definition

- $C \subseteq DB$ in subspace $S \subseteq D$ is a subspace cluster if:
  - **Objects in C are S-connected**: $\forall o, p \in C : \exists k : \forall i = 1, . . . , k - 1 : \exists\ q_i \in C : ||q_i - q_{i+1}||^S \leq \varepsilon \wedge q_1 = o, q_k = p$.
  - **C is maximal**, $\forall o, p \in DB$, o, p are S-connected $\Rightarrow (o \in C \Leftrightarrow p \in C)$.
  - **Minimum cluster size**: $|C| \geq minSize$.
  - **More dense than expected**
    $o \in C$ to be a cluster center : $\varphi^S(o) \geq F \cdot \alpha(S)$
  - **Not redundant**: $\neg \exists (C , S)$ subspace cluster with
    $C \subseteq C \wedge S \subset S \wedge |C | \geq r \cdot |C|$

Evaluation of clusters

- Quality
  - Avg. inverse entropy weighted by no. of objects per cluster
- Coverage
  - Percentage of objects in any subspace cluster.

Datasets :: Glass

- Number of Instances: 214
- Number of Attributes: ID + 9 + 1 class attribute
- Total no.of class labels:7 (1-7)
- Class distribution : 1 - 70, 2 - 76, 3 - 17, 4 - 0, 5 - 13, 6 - 9, 7 - 29

Results :

	DUSC(r = 0)		DUSC(r = 0.1)		SCHISM		SUBCLU	
	Exp	Paper	Exp	Paper	Exp	Paper	Exp	Paper
Attributes	M = 48 Eps = 0.25		M = 40 Eps = 0.25		xi=10 TAU=0.0045 U=0.05		M = 25 Eps = 10^5	
Time Taken	170ms		178ms		150 ms		30ms	
Quality	61	60	49	50	43	44	50	44
Coverage	88	87	94	93	97	99	100	100

Datasets :: Pendigits

- Number of Instances: 11472
- Number of Attributes: ID + 16 input + 1 class attribute
- Total no. of classes: 10(0-9)
- Class Distribution: 0 - 1143, 1 - 1143, 2 - 1144, 3 - 1055, 4 - 1144, 5 - 1055, 6 - 1056, 7 - 1142, 8 - 1055, 9 - 1055

Results :

	DUSC(r = 0)		DUSC(r = 0.1)		SCHISM		SUBCLU	
	Exp	Paper	Exp	Paper	Exp	Paper	Exp	Paper
Attributes	M = 100 Eps = 0.05		M = 80 Eps = 0.05		xi=10 TAU=0.0045 U=0.05		M = 80 Eps = 0.05	
Quality	80	86	75	81	53	77	43	58
Coverage	70	74	85	92	100	100	100	100

Conclusions

- As redundancy increases,
  - Coverage increases
  - Quality decreases slightly.
- Coverage is not 100%, indicating that DUSC algorithm is differentiating between noise and clusters in the subspaces of varying dimensions.
- Pendigits dataset has noise in it.
  - So , SUBCLU and SCHISM assign even this noise to clusters increasing value of coverage.

THANK YOU