Ranking Overlap and Outlier Points in Data using Soft Kernel Spectral Clustering

Raghvendra Mall, Rocco Langone and Johan A.K. Suykens

KU Leuven - ESAT/STADIUS
Kasteelpark Arenberg 10, B-3001 Leuven - Belgium
{raghvendra.mall,rocco.langone,johan.suykens}@esat.kuleuven.be

Abstract. Soft clustering algorithms can handle real-life datasets better as they capture the presence of inherent overlapping clusters. A soft kernel spectral clustering (SKSC) method proposed in [1] exploited the eigen-projections of the points to assign them different cluster membership probabilities. In this paper, we detect points in dense overlapping regions as overlap points. We also identify the outlier points by exploiting the eigen-projections. We then propose novel ranking techniques using structure and similarity properties in the eigen-space to rank these overlap and outlier points. By ranking the overlap and outlier points we provide an order for the most and least influential points in the dataset. We demonstrate the effectiveness of our ranking measures on several datasets.

1 Introduction

In the modern era where data can easily be collected from heterogeneous sources most real-life datasets have structure comprising of overlapping clusters. This has led to unsupervised learning models referred as Soft clustering methods [2, 3] which assign multiple cluster memberships to individual points in the data. These techniques can better deal with overlapping clusters and provide more insight about the data. For instance, when studying gene microarray datasets, genes that have more than one function by coding for proteins that participate in multiple metabolic pathways should belong to multiple overlapping clusters.

A kernel spectral clustering (KSC) method was proposed in [4] whose main advantage is its powerful out-of-sample extensions property which allows to generate eigen-projections for large scale data and infer their hard cluster affiliation. Recently, the KSC technique was extended to soft kernel spectral clustering (SKSC) method in [1]. The SKSC technique exploits the properties of the eigen-projections of the data to assign them multiple cluster memberships. This allows us to distinguish overlap points in dense overlapping regions from points which primarily belong to one cluster. Using the eigen-projections of the data it also possible to locate the outlier points.

The overlap points are more influential in the data as they have properties similar to multiple clusters in the data. These overlap points act as connectors between distinct clusters in the data. In the case of genes, the overlap genes are more important as they are part of multiple metabolic pathways and can provide more insight about the gene expressions. On the other hand, outlier points are the least influential points in the data and act as anomaly. They have properties which are dissimilar from most of the points in the data.

In this paper, we propose separate techniques to rank the overlap and outlier points in the dataset exploiting the structure and similarity properties of the

eigen-projections of these points. We develop an overlap score where higher rank for an overlap point is given by a lower overlap score and find that this overlap point is most similar to all the points in the data. We also develop an outlier score where a higher rank for an outlier point is given by higher score and find that this outlier point is least similar to all the points in the data.

2 Related Work

In information retrieval (IR) ranking is performed to provide an order in which the results corresponding to a particular query is displayed. A survey on various ranking techniques in information retrieval is provided in [5]. However, in IR ranking is based on similarity (i.e. in a classification setting) and overlap and outlier points are not generally considered while displaying the search results. There also exists a set of clustering algorithms which use ranking as a distance measure to obtain hard and soft clustering for datasets [6, 7]. To the best of our knowledge, this is the first approach where a soft clustering method is applied to obtain overlap and outlier points in the data and then these points are ranked to provide an ordering to the most and least influential points.

3 Identifying and Ranking Overlap & Outlier Points

We first briefly describe the SKSC [1] method. Given N_{tr} training points $\mathcal{D} = \{x_i\}_{i=1}^{N_{\text{tr}}}, x_i \in \mathbb{R}^{d_x}$ and k clusters, the KSC problem [4] can be stated as follows:

$$\min_{w^{(l)},e^{(l)},b_l} \frac{1}{2}\sum_{l=1}^{k-1} {w^{(l)}}^T w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1} \gamma_l {e^{(l)}}^T D_\Omega^{-1} e^{(l)} \tag{1}$$

$$\text{such that } e^{(l)} = \Phi w^{(l)} + b_l 1_{N_{\text{tr}}} \tag{2}$$

where $e^{(l)} = [e_1^{(l)}, \ldots, e_{N_{\text{tr}}}^{(l)}]^T$ are the projections vectors related to the N_{tr} training points, $D_\Omega^{-1} \in \mathbb{R}^{N_{\text{tr}} \times N_{\text{tr}}}$ is the inverse of the degree matrix associated to the kernel matrix Ω, Φ is the $N_{\text{tr}} \times n_h$ feature matrix $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_{N_{\text{tr}}})^T]$, $\varphi : \mathbb{R}^{d_x} \to \mathbb{R}^{n_h}$ is the mapping from input space (d_x) to a high-dimensional feature space (n_h), b_l are bias terms, and $\gamma_l \in \mathbb{R}^+$ are regularization constants. The corresponding dual is an eigen-decomposition problem which results in a dual solution given by $e^{(l)} = \Omega\alpha^{(l)} + b_l 1_{N_{\text{tr}}}$.

In SKSC method [1], KSC was used to first find a division of the data into k hard clusters. This clustering was then refined by re-calculating the prototypes in $e = [e^{(1)}, \ldots, e^{(k-1)}]$. In particular, given the projections for the training points $e_i, i = 1, \ldots, N_{tr}$ and the initial KSC hard cluster assignments (c_i), the new cluster prototypes $s_1, \ldots, s_p, \ldots, s_k$, $s_p \in \mathbb{R}^{k-1}$ became $s_p = \frac{1}{n_p}\sum_{i=1}^{n_p} e_i$ where n_p is the number of points assigned to cluster p during the initialization step by KSC. We then calculated the cosine distance (as proposed in [1]) between the i-th point projection and a prototype s_p as $d_{ip}^{\cos} = 1 - e_i^T s_p/(||e_i||_2||s_p||_2)$.

The probabilistic membership of point i to cluster p was expressed as:

$$m_i^{(p)} = \frac{\prod_{j \neq p} d_{ij}^{\cos}}{\sum_{l=1}^{k} \prod_{j \neq l} d_{ij}^{\cos}} \tag{3}$$

with $\sum_{l=1}^{k} m_i^{(l)} = 1$. This probability indicates certainty of SKSC membership. For model selection we use average membership strength (AMS) criterion [1].

3.1 Identifying Overlap & Outlier points

Using the membership probability, we devise a simple heuristic to detect overlap and outlier points. A point i is considered to lie in the overlap region between two or more clusters if its maximum soft membership $\max_p m_i^{(p)} < 0.5$.

One of the characteristics of an outlier point is that its similarity w.r.t. all the training points (N_{tr}) is close to 0. Using this property, a point is detected as outlier if its similarity with all the training points is small and its maximum soft cluster membership is higher than a threshold as it would have tendency to primarily belong to one cluster, i.e. $\sum_{i=1}^{N_{\text{tr}}} \Omega_{ij}^{\text{test}} < 10^{-2} N_{\text{tr}}$ and $\max_p m_i^{(p)} > 0.5$. We experimented with different values of this threshold and found that for values greater than 0.5 the set of outliers remain more or less consistent.

3.2 Ranking Score Functions

After identifying the overlap and outlier points, we create overlap set $\mathcal{D}_{ov} = \{x_i\}_{i=1}^{N_{ov}}$ and outlier set $\mathcal{D}_{out} = \{x_j\}_{j=1}^{N_{out}}$. Here N_{ov} and N_{out} represent the number of overlap and outlier points in the data respectively. We also create overlap projection set $\mathcal{E}_{ov} = \{e_i\}_{i=1}^{N_{ov}}$ and outlier projection set $\mathcal{E}_{out} = \{e_j\}_{j=1}^{N_{out}}$. We maintain hard and soft cluster memberships, $\mathcal{C}_{ov} = \{c_i\}_{i=1}^{N_{ov}}$, $c_i \in \mathbb{R}$ and $\mathcal{M}_{ov} = \{m_i\}_{i=1}^{N_{ov}}$, $m_i \in \mathbb{R}^k$ for overlap points. Similarly, we maintain hard and soft cluster memberships $\mathcal{C}_{out} = \{c_j\}_{j=1}^{N_{out}}$, $c_j \in \mathbb{R}$ and $\mathcal{M}_{out} = \{m_j\}_{j=1}^{N_{out}}$, $m_j \in \mathbb{R}^k$ for outlier points.

The overlap score consists of 3 components. The first component captures structural information and is given by: $\Delta_k(e_i) = \sum_{p=1}^{k} \|e_i - s_p\|_2 \times m_i^p$. It measures the distance of each overlap projection ($e_i \in \mathcal{E}_{ov}$) from a central projection of all the clusters giving more emphasis to the clusters to which it has higher probability of belonging ($m_i \in \mathcal{M}_{ov}$).

The second component comprises actual Euclidean distance of an overlap projection from all the projections weighted by extent of similarity. This component is inspired from an information retrieval aspect. In order to calculate this metric for all points with hard cluster membership p, we first estimate $\Delta_c(e_i, p) = [\|e_i - e_l\|_2 \; s.t. \; c_l = p]$. We then sort this vector and construct a weight vector $\omega_c(p) = [n_p, \ldots, 1]^\intercal$. More weights is given to smaller distance than to larger distance i.e. if an overlap projection is close to many projections in cluster p then it should have lower distance from that cluster. Finally, this component is estimated as $\Delta_{val}(e_i, p) = \frac{2 \times \Delta_c(e_i,p) \times \omega_c(p)}{n_p \times (n_p+1)}$. The overall weighted distance for the i^{th} overlap projection (e_i) is: $\Delta_\omega(e_i) = \sum_{p=1}^{k} \Delta_{val}(e_i, p) \times m_i^p$.

The third component comprises of the similarity of an overlap point $x_i \in \mathcal{D}_{ov}$ from all the points in the dataset in terms of the kernel matrix Ω. An overlap point has high similarity value w.r.t. most of the points in the data. This helps to distinguish an influential overlap point from a mis-categorized outlier point which has low similarity value w.r.t. all points in the data. This component is represented as: $S_{val}(x_i) = \sum_{l=1}^{N} \Omega_{il}$.

We then combine these 3 components to devise a scoring scheme which gives higher rank to overlap points which are part of dense overlapping regions. The overlap score for the i^{th} point in the overlap set $\mathcal{D}_{ov}$ is calculated as:

$$sc_{ov}(i) = \frac{\Delta_k(e_i) \times \Delta_\omega(e_i)}{S_{val}(x_i)}. \quad (4)$$

In the score function the distance terms are kept in the numerator and the similarity term is used as the denominator. We want to minimize the distance terms and maximize the similarity term for an overlap point in the score function. Thus, smaller values of $sc_{ov}(\cdot)$ give higher rank indicating these points have characteristics similar to points in multiple clusters and are more influential.

We use the property that the similarity of an outlier point w.r.t. all the points in the data is extremely small i.e. $\Omega_{ij} \approx 0$, $i = 1, \ldots, N$,$j = 1, \ldots, N_{out}$, $x_i \in \mathcal{D}$ and $x_j \in \mathcal{D}_{out}$. Using this property and dual solution of KSC, we conclude that the eigen-projection of an outlier point can be given as: $e_j \approx b$, where $b = [b_1, \ldots, b_{k-1}]^\intercal$. In the ideal case, an outlier will have 0 similarity w.r.t. all the points in the data and its eigen-projection will be exactly $= b$. Using this notion we define a distance measure for outlier points as: $\Delta_{out}(e_j) = \sum_{p=1}^{k}(\|e_j - s_p\|_2 - \|b - s_p\|_2) \times m_j^p$.

Here $e_j \in \mathcal{E}_{out}$ and $m_j \in \mathcal{M}_{out}$. This metric evaluates the distance of an outlier eigen-projection (e_j) from the cluster prototypes (s_p) and calculates the same for the bias vector (b). It gives more weight (m_j^p) to the difference in distance for the cluster to which this outlier actually belongs. This metric is more robust than a simple Euclidean distance ($\|e_j - b\|_2$) as it includes the influence of the soft clustering memberships for outlier points. The smaller the value of this distance measure ($\Delta_{out}(e_j)$), the lower the significance of that outlier. However, these values can be quite small (≈ 0) at times and difficult to interpret. Hence, we define the $sc_{out}(\cdot)$ function as:

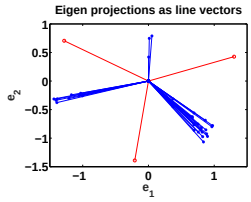
$$sc_{out}(e_j) = 1 - \Delta_{out}(e_j). \quad (5)$$

Larger the value of this $sc_{out}(\cdot)$ function for an outlier higher the rank, since the similarity of this outlier w.r.t. any point in the dataset is low. Figure 1 shows the location of the overlap and outlier points detected by SKSC method in the input space and eigen-space for a synthetic 3 overlapping Gaussians dataset.

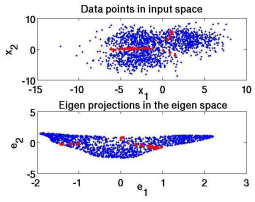
4 Experiments

We conducted experiments on 10 datasets obtained from `http://cs.joensuu.fi/sipu/datasets/`. Figure 2 shows the model selection procedure, the overlap, outlier points and the clustering generalization for A1 and Mopsi Finland datasets. Table 1 shows number of overlap and outlier points detected by SKSC method in these datasets. Outlier points are ranked based on the proposed sc_{out} for 3 datasets in Table 2. The higher the sc_{out} value lesser the similarity of that point w.r.t. any point in the dataset. It allows us to easily identify least influential points in the dataset.

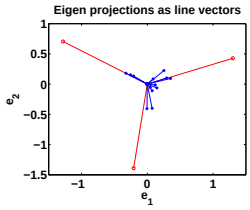
We compare our proposed sc_{ov} based ranking with distance based ranking technique (D-Rank or D-R) [6] and information retrieval (similarity) based ranking technique (IR-Rank or IR-R) [5] as shown in Table 3 for A1, Mopsi Finland and Mopsi Joensuu datasets. We calculate the Kendall τ ranking correlation between the ranking order of proposed method with D-Rank and IR-Rank. For A1, Mopsi Finland and Mopsi Joensuu datasets the correlation values are (-0.1,



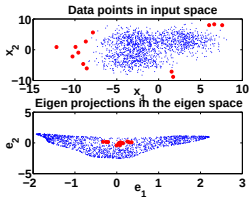
(a) Red lines represent cluster prototypes s_i and blue lines show overlap projections.



(b) Red colored points represent the overlap points in input and eigen-space.

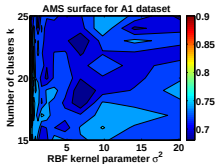


(c) Red lines represent cluster prototypes s_i and blue lines show outlier projections.

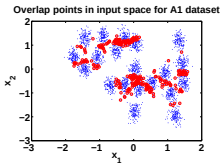


(d) Red colored points represent the outlier points in input and eigen-space.

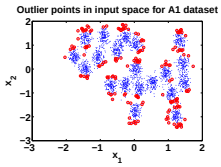
Fig. 1: Structure of the overlap and outlier points in the eigen and input space for a synthetic 3 overlapping 2-dimensional Gaussians.



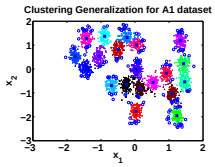
(a) AMS surface for A1 dataset.



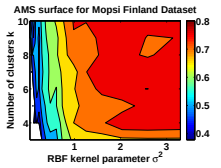
(b) Red points show overlap points.



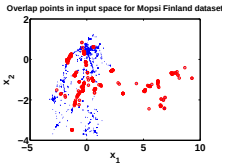
(c) Red points show outlier points.



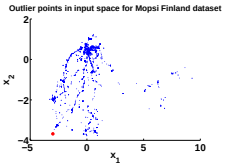
(d) Generalization results.



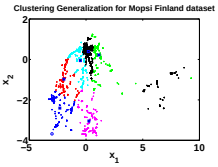
(e) AMS surface for Mopsi Finland data.



(f) Red points show overlap points.



(g) Red point represent outlier point.



(h) Generalization results.

Fig. 2: Tuning of SKSC algorithm, detection of overlap and outlier points and cluster generalization for 2 datasets obtained from `http://cs.joensuu.fi/sipu/datasets/`.

-0.005), (0.123, 0.218), (0.355 and 0.45) w.r.t. D-Rank and IR-Rank respectively. In general we observe low correlation between the rankings.

We ran our proposed approach on a NIPS dataset comprising $1,500$ papers available at `https://archive.ics.uci.edu/`. No outlier documents and 188 overlap documents were detected using SKSC [1]. The most influential paper was "Adaptive Development of Connectionist Decoders for Complex Error-Correcting Codes (ECC)". ECC is a popular approach to handle multi-class

problems for many supervised learning techniques making it highly influential.

Dataset	N	d_x	k	N_{ov}	N_{out}
A1	3,000	2	20	281	165
Aggregation	788	2	5	32	-
Europe	169,308	2	2	-	81
Iris	150	4	3	3	-
Mopsi Finland	13,467	2	6	510	1
Mopsi Joensuu	6,014	2	4	35	31
R15	600	2	15	11	5
Seeds	210	7	3	5	-
3 Gaussians	1,500	2	3	38	13
Wine	178	13	3	6	-

Table 1: N_{ov} and N_{out} represent number of overlap and outlier points and '-' means that no overlap or no outlier point.

A1 dataset		Europe dataset		Mopsi Joensuu	
Point Id	sc_{out}	**Point Id**	sc_{out}	**Point Id**	sc_{out}
1. 864	0.999	1. 163927	0.947	1. 5732	1
2. 2951	0.996	2. 162749	0.929	2. 5734	0.998
3. 2734	0.981	3. 956360	0.855	3. 5728	0.998
4. 2042	0.875	4. 157332	0.827	4. 5731	0.996
5. 1263	0.710	5. 151013	0.788	5. 1951	0.867
6. 1420	0.579	6. 735160	0.785	6. 1146	0.865
7. 2935	0.574	7. 126906	0.784	7. 1949	0.865
8. 993	0.565	8. 735140	0.782	8. 1647	0.864
9. 1983	0.518	9. 144385	0.781	9. 1652	0.864
10. 2006	0.482	10. 95557	0.781	10. 5772	0.853

Table 2: Outlier ranking results showing the least influential outlier points in order produced by the proposed sc_{out} for A1, Europe and Mopsi Joensuu dataset.

A1 dataset				**Mopsi Finland dataset**				**Mopsi Joensuu dataset**			
Point Id	sc_{ov}	**D-R**	**IR-R**	**Point Id**	sc_{ov}	**D-R**	**IR-R**	**Point Id**	sc_{ov}	**D-R**	**IR-R**
1. 2092	47.268	160	99	1. 3172	53.478	240	65	1. 5051	27.273	1	1
2. 2000	47.895	153	83	2. 3174	53.48	239	66	2. 1183	28.597	10	5
3. 2086	48.34	150	39	3. 3078	53.483	235	67	3. 3911	28.772	12	4
4. 2055	49.625	146	57	4. 3105	53.484	238	69	4. 3910	28.777	13	3
5. 2093	49.999	157	30	5. 2662	53.485	236	68	5. 1184	28.854	14	2
6. 2011	50.043	166	24	6. 3254	53.505	234	70	6. 1721	29.068	17	6
7. 2032	52.509	170	21	7. 3462	53.522	237	71	7. 1978	67.001	3	9
8. 2069	53.924	149	92	8. 458	69.394	162	1	8. 1634	67.009	4	10

Table 3: Ranking results showing the top 8-ranked overlap/influential points produced by proposed sc_{ov} for A1, Mopsi Finland and Mopsi Joensuu datasets and its comparison with D-Rank (D-R) and IR-Rank (IR-R).

5 Conclusion

We proposed a technique to identify and rank overlap and outlier points in data by exploiting the structure and similarity property of these points in eigen-space using the SKSC method. In future, we would like to quantify the relevance of the proposed ranking scheme w.r.t. other ranking techniques.

Acknowledgments: The work is supported by Research Council KUL, ERC AdG A-DATA-DRIVE-B, GOA/10/09MaNet, CoE EF/05/006, FWO G.0588.09, G.0377.12, SBO POM, IUAP P6/04 DYSCO.

References

[1] R. Langone, R. Mall, and J.A. K. Suykens. Soft kernel spectral clustering. In *Proc. of the IJCNN 2013*, pages 1028 – 1035, 2013.

[2] H.C. Huang, Y.Y. Chuang, and C.S. Chen. Multiple kernel fuzzy clustering. *Fuzzy Systems, IEEE Transactions on*, 20(1):120–134, February 2012.

[3] K. Yu, S. Yu, and V. Tresp. Soft clustering on graphs. In *in Advances in Neural Information Processing Systems (NIPS)*, page 05, 2005.

[4] C. Alzate and J. A. K. Suykens. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):335–347, February 2010.

[5] J. Datta and P. Bhattacharya. Ranking in information retrieval. Technical report, Indian Institute of Technology, Bombay, 2010.

[6] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogenous networks with star network schema. In *Proc of KDD*, 2009.

[7] S. Rovetta, F. Masulli, and M. Filippone. Soft rank clustering. *Neural Nets, Lecture Notes in CS*, 3931:207–213, 2006.