Kernel Spectral Document Clustering Using Unsupervised Precision-Recall Metrics

Raghvendra Mall
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001, Leuven, Belgium
Email: raghvendra.mall@esat.kuleuven.be

Johan A.K. Suykens
KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001, Leuven, Belgium
Email: johan.suykens@esat.kuleuven.be

***Abstract*—Kernel Spectral Clustering (KSC) solves a weighted kernel principal component analysis problem in a primal-dual optimization framework. The KSC model is built on a small subset of data using a proper training, model selection and a test phase. The clustering model is obtained using the dual solution of the problem and has a powerful out-of-sample extensions property which allows cluster affiliation for previously unseen data points. In the model selection phase, we estimate the appropriate number of clusters using a metric that evaluates the quality of the clusters. Traditional quality indices like inertia, Davies-Bouldin (DB) index and silhouette (SIL) are known to be method-dependent and not perform well in case of complex heterogeneous data like textual data. In this paper, we utilize the quality evaluation techniques based on an unsupervised version of Precision, Recall and F-measure proposed in [1] to come up with a new kernel spectral document clustering (KSDC) model which generates homogeneous clusters of documents. We compare the quality of the clusters obtained by the proposed KSDC technique with k-means and neural gas algorithm, which are more oriented towards these metrics, on several real world textual data.**

I. Introduction

Clustering algorithms are widely used tools in fields like data mining, machine learning, graph compression, text clustering, community detection and many other tasks. The aim of clustering is to divide the data into natural groups present in a given dataset. Clusters are defined such that the data present within the group are more similar to each other in comparison to the data between clusters. In this paper, we focus on the specific problem of document clustering in complex heterogeneous textual data. The problem of document clustering finds its application in a variety of domains including document organization and browsing [2], corpus summarization and document classification [3], [4]. A brief survey of various text clustering algorithms is provided in [5], [6].

Spectral clustering methods [7], [8], and [9] work in the latent space. They project the input data to an eigenspace based on the Laplacian matrix and then perform k-means to obtain localized clusters in the eigenspace. A non-negative matrix factorization (NMF) based method was utilized for document clustering in [10]. However, a drawback of traditional spectral clustering methods is that they are computationally expensive ($O(N^3)$) and memory inefficient $O(N^2)$, where N is the number of documents in the text data.

A new Kernel Spectral Clustering (KSC) algorithm based on weighted kernel PCA formulation was proposed in [11]. The method was based on building the clustering model on a small subset of data in a primal-dual optimization framework. The clustering model has a powerful out-of-sample extension property which allows to infer cluster affiliation for unseen data. The KSC methodology has been extensively applied for task of data clustering [11], data classification [12], [13] and community detection [14], [15], [16] and [17] in large scale networks.

The quality of the clusters are considered good if they possess low intra-cluster distances as compared to their inter-cluster distances. However, it was shown by the authors in [18] that the distance based indices like Davies-Bouldin (DB) and Silhouette (SIL) [19] are often strongly biased and highly dependent on the clustering method. It was also pointed out by the authors in [20] that the experiments on these indices in the literature are often performed on unrealistic test corpora constituting of low-dimensional data. It was shown in [21] that the aforementioned indices are often unable to identify an optimal clustering model when the dataset has complex structure that must be represented in a both high-dimensional and sparse description space as it is often the case with textual data. In [1], the authors proposed clustering quality metric based on Precision, Recall and F-measure using evaluation principles from information retrieval. A major advantage of these quality metrics is that they are independent of clustering methods and their operating mode as shown in [1]. These quality metrics are more favourable for text data since they try to locate homogeneous clusters looking at word content of individual documents.

Taking into consideration the success of kernel spectral clustering (KSC) for large scale networks (which are both sparse and high-dimensional), in this paper, we perform kernel spectral document clustering (KSDC) using the quality metrics proposed in [1] during model selection to obtain the optimal k. We show the effectiveness of KSDC methodology by comparing its results with k-means [22], [23] and neural gas (NG) [24] algorithms, which have been shown to perform well in combination with the precision and recall based metrics [1], on several real world textual data.

In Section II we provide a brief description of the kernel spectral clustering (KSC) technique. In Section III, we outline the primary contribution of this paper. Section IV details the experimental setting and analysis of the results obtained. We conclude the paper in section V.

II. Kernel Spectral Clustering

We first explain the data preparation step of textual corpus and then provide a brief description of the kernel spectral clustering methodology [11]. We obtained the text corpora from UCI repository [24] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words, http://www.cs.umb.edu/~smimarog/textmining/datasets/ and http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html. In the case of the text datasets obtained from UCI repository, we perform TF-IDF [25] weighting of words in documents whereas in the latter cases the TF-IDF weighting is already performed. The TF-IDF weighting scheme is a more enhanced representation where words in documents are weighted based on the frequencies of the individual words in the document as well as frequencies of words in the entire text corpora. Now a document $x_i \in \mathbb{R}^d$ is represented as a bag-of-words where d is the total number of words with non-zero weights in the text corpora. Thus, each x_i is high-dimensional and sparse since not all words are present in each document. We now briefly describe the kernel spectral clustering [11] model.

A. Primal-Dual Weighted Kernel PCA framework

Given a dataset $\mathcal{D} = \{x_i\}_{i=1}^{N_{tr}}$, $x_i \in \mathbb{R}^d$ and the number of clusters k, the primal problem of the kernel spectral clustering (KSC) via weighted kernel PCA is formulated as follows [11]:

$$\begin{aligned} \min_{w^{(l)},e^{(l)},b_l} \quad & \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)\intercal} w^{(l)} - \frac{1}{2N_{tr}}\sum_{l=1}^{k-1}\gamma_l e^{(l)\intercal} D_\Omega^{-1} e^{(l)} \\ \text{such that} \quad & e^{(l)} = \Phi w^{(l)} + b_l 1_{N_{tr}}, l = 1,\ldots,k-1, \end{aligned} \tag{1}$$

where $e^{(l)} = [e_1^{(l)},\ldots,e_{N_{tr}}^{(l)}]^\intercal$ are the projections onto the eigenspace, $l = 1,\ldots,k-1$ indicates the number of score variables required to encode the k clusters, $D_\Omega^{-1} \in \mathbb{R}^{N_{tr}\times N_{tr}}$ is the inverse of the degree matrix associated to the kernel matrix Ω. Φ is the $N_{tr} \times n_h$ feature matrix, $\Phi = [\phi(x_1)^\intercal;\ldots;\phi(x_{N_{tr}})^\intercal]$ and $\gamma_l \in \mathbb{R}^+$ are the regularization constants. We note that $N_{tr} \ll N$ i.e. the number of points in the training set is much less than the total number of data points in the dataset. Each element of Ω, denoted as $\Omega_{ij} = K(x_i,x_j) = \phi(x_i)^\intercal\phi(x_j)$ is obtained by calculating the cosine similarity between x_i and x_j which has been shown to be an effective similarity measure in case of text corpora [26]. Thus, $\Omega_{ij} = \frac{x_i^\intercal x_j}{\|x_i\|\|x_j\|}$ and can be calculated efficiently using notions of set unions and intersections. This corresponds to using a normalized linear kernel function $K(x,z) = \frac{x^\intercal z}{\|x\|\|z\|}$ [27]. The clustering model in the primal is then represented by:

$$e_i^{(l)} = w^{(l)\intercal}\phi(x_i) + b_l, i = 1,\ldots,N_{tr}, \tag{2}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the mapping to a high-dimensional feature space n_h, b_l are the bias terms, $l = 1,\ldots,k$-1. However, for textual data, we can utilize the explicit expression of the underlying feature map and $n_h = d$. The projections $e_i^{(l)}$ represent the latent variables of a set of k-1 binary cluster indicators given by $\mathrm{sign}(e_i^{(l)})$ which can be combined with the final groups using an encoding/decoding scheme. The decoding consists of comparing the binarized projections w.r.t. codewords in the codebook and assigning cluster membership based on minimal Hamming distance. The Lagrangian of the problem 1 is given by:

$$\begin{aligned} \mathcal{L}(w^{(l)},e^{(l)},b_l;\alpha^{(l)}) = \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)\intercal}w^{(l)} - \frac{1}{2N_{tr}}\sum_{l=1}^{k-1}\gamma_l e^{(l)\intercal} D_\Omega{}^{-1} e^{(l)} \\ - \sum_{l=1}^{k-1}\alpha^{(l)\intercal}(e^{(l)} - \Phi w^{(l)} - b_l 1_{N_{tr}}), \end{aligned} \tag{3}$$

with KKT optimality conditions:

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w^{(l)}} = 0 \to w^{(l)} = \Phi^\intercal\alpha^{(l)}, \\ \frac{\partial\mathcal{L}}{\partial e^{(l)}} = 0 \to \alpha^{(l)} = \frac{\gamma_l}{N_{tr}} D_\Omega{}^{-1} e^{(l)}, \\ \frac{\partial\mathcal{L}}{\partial b_{(l)}} = 0 \to 1_{N_{tr}}^\intercal\alpha^{(l)} = 0, \\ \frac{\partial\mathcal{L}}{\partial\alpha^{(l)}} = 0 \to e^{(l)} = \Phi w^{(l)} + b_l 1_{N_{tr}}, \end{cases}$$

for $l = 1,\ldots,k-1$. The bias term becomes:

$$b_l = -\frac{1}{1_{N_{tr}}^\intercal D_\Omega^{-1} 1_{N_{tr}}} 1_{N_{tr}}^\intercal D_\Omega^{-1}\Omega\alpha^{(l)}, l = 1,\ldots,k-1$$

The dual problem corresponding to this primal formulation is:

$$D_\Omega^{-1} M_D \Omega\alpha^{(l)} = \lambda_l\alpha^{(l)}, \tag{4}$$

where M_D is the centering matrix which is defined as $M_D = \mathrm{I}_{N_{tr}} - (\frac{(1_{N_{tr}} 1_{N_{tr}}^\intercal D_\Omega^{-1})}{1_{N_{tr}}^\intercal D_\Omega^{-1} 1_{N_{tr}}})$. The $\alpha^{(l)}$ are the dual variables and the positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ plays the role of similarity function. This dual problem is closely related to the random walk model as shown in [11].

B. Out-of-Sample Extensions Model

The projections $e^{(l)}$ define the cluster indicators for the training data. In the case of an unseen data point (document) x, the predictive model becomes:

$$\hat{e}^{(l)}(x) = \sum_{i=1}^{N_{tr}} \alpha_i^{(l)} K(x,x_i) + b_l. \tag{5}$$

This out-of-sample extension property allows KSC to be formulated in a learning framework with training, model selection and test stage for better generalization. The model selection/validation stage is used to obtain the optimal model parameter i.e. the number of clusters k in the text dataset.

III. Proposed Subset Selection & Model Selection Criterion for Document Clustering

A. Subset Selection

From the KSC formulation it can be observed that it is essential to have a representative subset of the large scale text corpora as training and validation set in order to capture the optimal number of clusters k in the dataset. In order to achieve this objective we first convert the text corpus comprising documents into a weighted r-NN (r Nearest Neighbour) graph. In the r-NN graph, for each document in the text data, we locate the top r similar documents using the cosine similarity metric and weight of the edges represent the extent of similarity. We then perform a fast and unique representative subset (FURS) selection technique [15] to obtain the set of documents which are part of different dense regions in the r-NN graph as training

and validation set. This technique was recently proposed in [28] and was shown to be scalable on a distributed environment for large scale datasets. We use $r = 100$ in our experiments as it was shown in [28] that the clustering performance is best for this value of r. We select the size of training and validation set as $\min(\frac{15N}{100}, 5000)$ based on memory constraints of a desktop computer with 8Gb Ram.

B. Model Selection

The original KSC formulation in [11] works well assuming piece-wise constant eigenvectors and exploits the line structure of the validation projections in the eigenspace. In Table I, we provide the strengths of various clustering quality metrics which are based on the structure of the eigen-projections.

TABLE I: Comparison of different model selection criteria for KSC exploiting the projection structures in the eigenspace.

Criterion	Strengths
BLF [11]	Works best in case of well-separated, non-overlapping clusters in input space with clear line-structure in eigenspace
BAF [14]	Works well in case of ≥ 3 overlapping clusters by exploiting angular similarity of eigen-projections to mean projection vector in each cluster.
AMS [29]	Based on the same principle as **BAF** but able to identify $k = 2$ clusters.
SIL [19]	Based on principle of intra and inter cluster euclidean distance in the eigenspace.

The aforementioned criteria work well for datasets where the eigen-projections have a specific structure and there exists a small number of large-sized clusters. However, in case of textual data, we generally observe a large number of clusters with fewer documents belonging to each category [1]. Thus, for text datasets we utilize unsupervised Precision, Recall and F-measure criterion proposed in [1]. A major advantage of this technique is the ability to produce homogeneous clusters and distinguish them from heterogeneous or degenerated ones.

Once the cluster affiliation for a given value of k is known using the codebook, encoding and decoding scheme for the validation set of documents, we then evaluate the quality of clusters using these unsupervised information retrieval metrics. Given the validation set of documents $\mathcal{V}$, we modify the Recall (Rec) and Precision ($Prec$) indices for a particular property (word) p of the cluster c introduced in [1] and express it as:

$$Rec_c(p) = \frac{\sum_{d \in c} W_d^p}{\sum_{c' \in \mathcal{C}} \sum_{d \in c'} W_d^p},$$
$$Prec_c(p) = \frac{\sum_{d \in c} W_d^p}{\max_{p'}(\sum_{d \in c} W_d^{p'})}$$

where W_x^p represents the weight of property p in document x and $\mathcal{C}$ represent the set of clusters in the validation set. Then, to estimate overall clustering quality, we utilize the average Macro-Recall (R_M) and Macro-Precision (P_M) defined as:

$$R_M = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} Rec_c(p), \tag{6}$$

$$P_M = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} Prec_c(p) \tag{7}$$

where S_c is the set of peculiar properties of the cluster c and $\bar{C}$ represents the peculiar set of clusters extracted from the set $\mathcal{C}$ as defined in [1]. Macro-Recall and Macro-Precision properties are average values of Recall and Precision for each cluster. They have antagonistic behaviour and optimum number of clusters k can be detected when both these values are high. So, we define a Macro F-measure (F_M) as:

$$F_M = \frac{2 \times R_M \times P_M}{R_M + P_M}$$

and try to maximize the same. F_M takes values between $[0, 1]$.

A drawback of this F_M measure is that it does not permit to detect degenerated clustering results when there is presence of small number of heterogeneous clusters as shown in [1]. To overcome this issue, the authors in [1] proposed property-oriented indices of Micro-Recall (R_m) and Micro-Precision (P_m) which are defined as:

$$R_m = \frac{1}{d} \sum_{c \in \bar{C}, p \in S_c} Rec_c(p), \tag{8}$$

$$P_m = \frac{1}{d} \sum_{c \in \bar{C}, p \in S_c} Prec_c(p). \tag{9}$$

In order to identify heterogeneous clusters of large size, the authors in [1] modified the Micro-Precision to operate cumulatively. The idea was to give a minor weights to large clusters which are most likely to re-partition and significantly lower their contribution in the quality metric. The Cumulative Micro-Precision (CP_m) is defined as:

$$CP_m = \frac{\sum_{i=|c_{inf}|,|c_{sup}|} \frac{1}{|C_{i+}^2|} \sum_{c \in C_{i+}, p \in S_c} Prec_c(p)}{\sum_{i=|c_{inf}|,|c_{sup}|} \frac{1}{C_{i+}}}, \tag{10}$$

where C_{i+} represents the subset of clusters of $\mathcal{C}$ for which associated data is greater than i, and: $inf = \text{argmin}_{c_i \in \mathcal{C}} |c_i|$, $sup = \text{argmax}_{c_i \in \mathcal{C}} |c_i|$.

We then define a Micro F-measure (F_m) as:

$$F_m = \frac{2 \times R_m \times CP_m}{R_m + CP_m}.$$

This metric takes value between $[0, 1]$ and optimizes the clustering quality for a particular k w.r.t. both R_m and CP_m.

The goal of model selection in KSC is to identify the optimal number of clusters k in the text data such that the quality of obtained clusters is maximum i.e. we obtain homogeneous clusters and are able to distinguish them from degenerate ones. Thus, we need to optimize w.r.t. both F_M and F_m. While the Macro F-measure is largest for smaller values of k and decreases as we increase the value of k, the Micro F-measure is smaller for lower values of k and increases as we find large number of small homogeneous clusters. So, we devise a new metric which minimizes the absolute difference between the values of F_M and F_m, thereby retaining both macro and micro qualities. The metric is defined as:

$$Q = 1 - |F_M - F_m|. \tag{11}$$

This metric is bounded to take values between $[0, 1]$ and higher values of Q represent better quality clusters. Figure 1 showcases the model selection procedure for R8 of

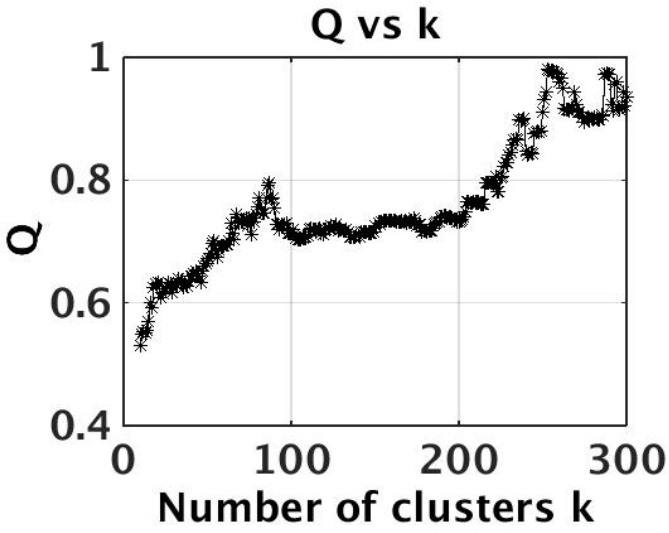


Fig. 1: Optimal number of clusters detected by Q metric is 253 for which the $F_M = 0.339$ and $F_m = 0.319$ for R8 dataset. For $k > 300$ the value of Q starts to decrease gradually.

Reuters 21578 dataset obtained from http://www.cs.umb.edu/~smimarog/textmining/datasets/.

Algorithm 1 summarizes the kernel spectral document clustering (KSDC) technique using modified version of precision and recall metrics.

IV. Experiments

A. Experimental Data

We conducted experiments on 10 real world text corpus obtained from various text mining sites including https://archive.ics.uci.edu/ml/datasets/Bag+of+Words, http://www.cs.umb.edu/~smimarog/textmining/datasets/ and http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html. We provide a brief description of these textual data in Table II.

We compare our proposed KSDC approach with traditional k-means [23] and neural gas (NG) [24] algorithms using cosine distance [26] as a distance measure. It was shown by the authors in [1] that these algorithms are more favorable for the unsupervised precision-recall based metrics. As mentioned earlier, we create a r-NN graph from the documents ($r = 100$ in our experiments) and then perform the fast and unique representative subset (FURS [15]) selection to obtain the validation set of documents. This subset of documents are representative set of all the documents present in the text corpus and retain the inherent cluster structure. In order to have a fair comparison the search for optimal k is performed on this validation set for a range between $[10, 300]$ using the proposed Q quality metric for all the clustering algorithms. Figure 2 highlights the tuning procedure for KSDC, k-means and NG algorithms in case of Nips, Kos, Classic4 and Reuters text datasets.

After obtaining the optimal k for KSDC method, we perform the out-of-sample extensions to obtain cluster affiliation for previously unseen text data. We use the entire text corpus as test data in our experiments. In case of KSDC, we need to perform the entire clustering procedure only once as the training and validation set are unique and obtained by a deterministic procedure [15] and the resulting clustering model

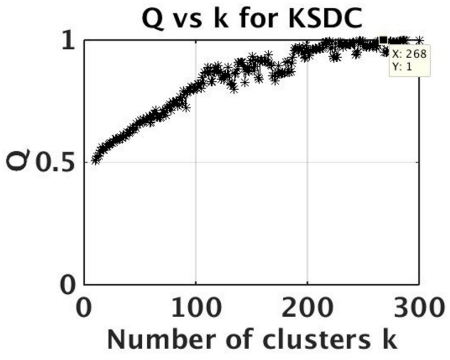
Algorithm 1: Kernel Spectral Document Clustering (KSDC)

Data: Given a document dataset $\mathcal{T} = \{x_i\}_{i=1}^{N}$.
Result: The partitions of these documents i.e. categorize the dataset into k clusters.

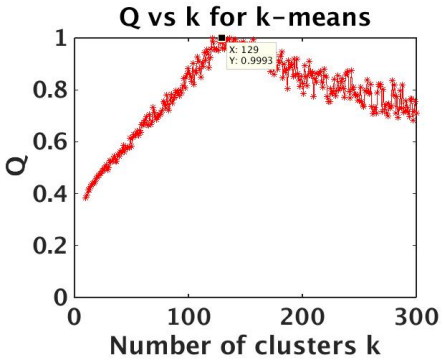
1 Convert the documents into a bag-of-words sparse representation using TF-IDF [25] weighting scheme.
2 Generate a r-NN graph between the documents using cosine similarity metric.
3 Select the training set $\mathcal{D}$ (i.e. maximum size $= 5,000$ documents) using FURS [15].
4 Select the validation set (same size as $|\mathcal{D}|$) after the removing documents from the set $\mathcal{D}$ corresponding to the training data using FURS [15].
5 Calculate the kernel matrix Ω by applying cosine similarity operations on sparse format data $\forall i, j$ $x_i, x_j \in \mathcal{D}$.
6 Perform eigen-decomposition of Ω to obtain the model i.e. $\alpha^{(l)}, b_l$.
7 Use out-of-sample extension property to obtain the projection values for the validation set i.e. $e^{(l)}_{valid}$.
8 **foreach** k **do**
9 Use codebook, encoding and decoding scheme as proposed in [11] to obtain cluster memberships.
10 **Calculate the unsupervised quality metrics like Macro F-measure (F_M) and Micro F-measure (F_m).**
11 **Estimate the value of $Q = 1 - |F_M - F_m|$ for each k.**
`// These step are performed by our proposed approach.`
12 **end**
13 Select the k corresponding to which the Q is maximum.
14 After estimating k, use the out-of-sample property to assign clusters to unseen test data. (We use the entire corpus as test data).
15 Estimate the F_M and F_m for the entire corpus.

TABLE II: Text data used in the experiments. Here N represents number of documents, d represents the number of words obtained after performing TF-IDF weighting scheme and nnz represents the number of non-zero counts in the bag-of-words.

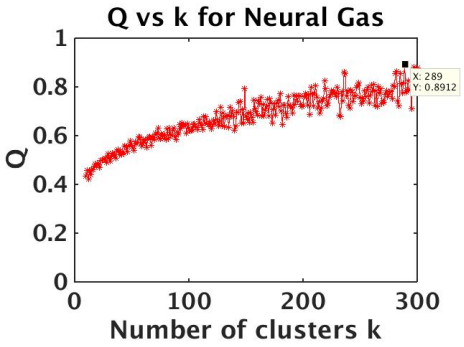
Dataset	N	d	nnz
Nips	1,500	115,60	217,900
Kos	3,430	5,631	58,146
R8	6,743	6,737	77,582
Classic4	7,095	5,896	247,158
Reuters	8,293	18,933	389,455
TDT2	9,394	36,771	1,224,135
TDT2all	10,212	36,771	1,323,869
20NewsGroups	18,774	61,188	2,435,219
Enron	36,439	24,311	1,009,003
NYTimes	300,000	82,358	90,180,242



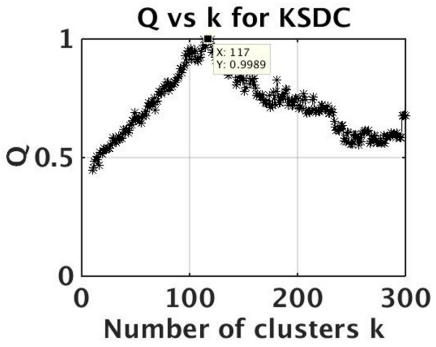
(a) $k = 268$ for KSDC (Nips)



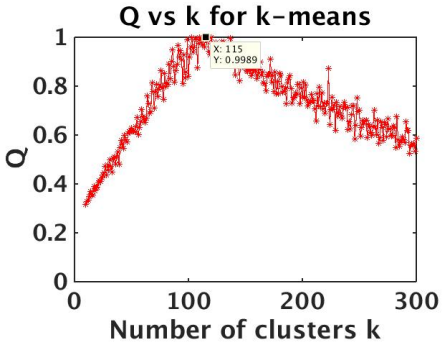
(b) $k = 129$ for k-means (Nips)



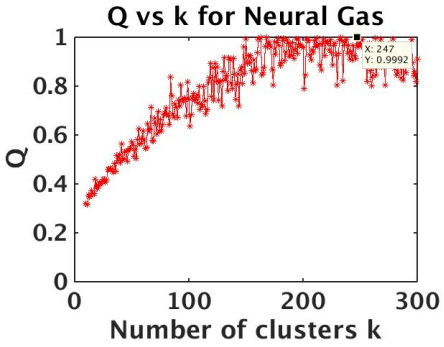
(c) $k = 289$ for NG (Nips)



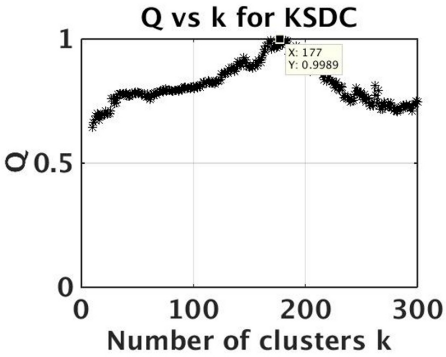
(d) $k = 117$ for KSDC (Kos)



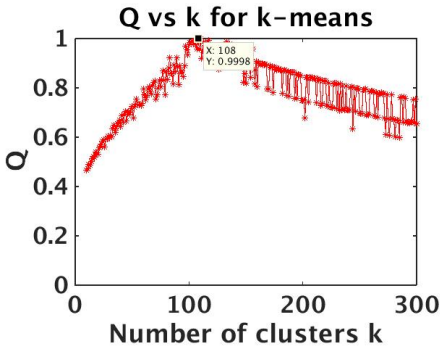
(e) $k = 115$ for k-means (Kos)



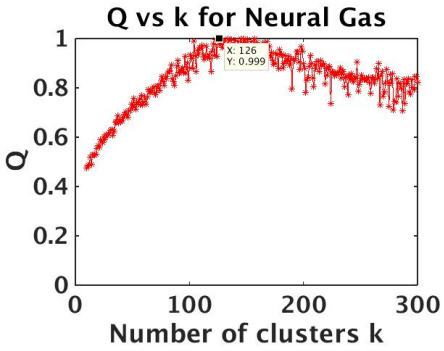
(f) $k = 247$ for NG (Kos)



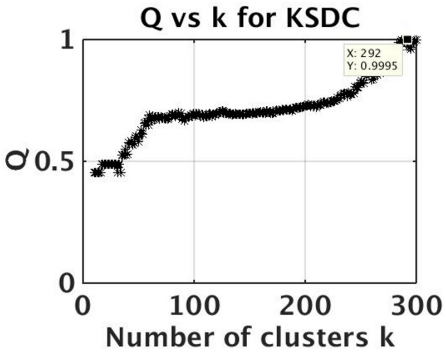
(g) $k = 177$ for KSDC (Classic4)



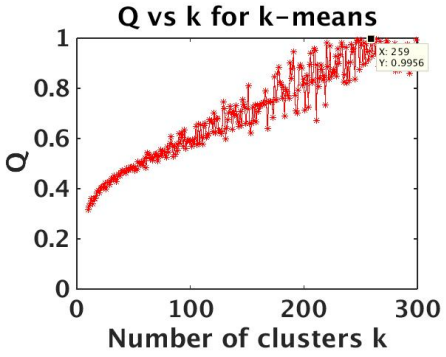
(h) $k = 108$ for k-means (Classic4)



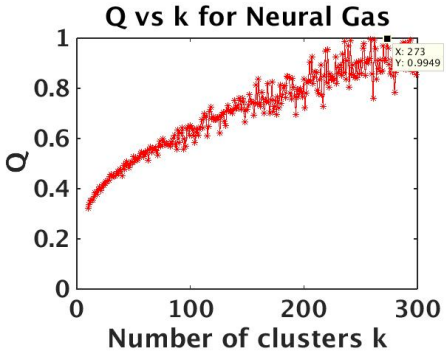
(i) $k = 126$ for NG (Classic4)



(j) $k = 292$ for KSDC (Reuters)



(k) $k = 259$ for k-means (Reuters)



(l) $k = 273$ for NG (Reuters)

Fig. 2: Comparison of proposed Q quality metric for KSDC, k-means and NG algorithms on Nips, Kos, Classic4 and Reuters datasets. For fair comparison the optimal k search using the Q quality metric is performed on the validation set for all the methods. The final clustering quality of the entire corpus can be different from the validation set. We observe that the trends obtained from KSDC are relatively smoother in comparison to k-means and NG due to their dependence on initial randomization.

TABLE III: Comparison of clustering quality using F_M, F_m and Q metrics for KSDC, k-means and NG algorithms. The highlighted results represent the maximum value for that metric and need not necessarily be the best.

Dataset	KSDC				k-means				Neural Gas (NG)			
	k	F_M	F_m	Q	k	F_M	F_m	Q	k	F_M	F_m	Q
Nips	268	0.274	**0.328**	**0.946**	129	**0.348**	0.222	0.874	289	0.292	0.075	0.783
Kos	117	0.386	**0.366**	**0.980**	115	**0.446**	0.300	0.854	247	0.392	0.295	0.903
R8	253	0.338	0.342	**0.996**	146	**0.542**	**0.408**	0.886	102	0.497	0.117	0.620
Classic4	177	0.241	**0.203**	**0.962**	108	0.289	0.130	0.841	126	**0.293**	0.143	0.85
Reuters	292	0.286	0.273	**0.987**	259	**0.332**	**0.275**	0.943	273	0.287	0.133	0.846
TDT2	234	0.230	**0.135**	**0.905**	227	**0.265**	0.123	0.858	293	0.203	0.085	0.882
TDT2all	254	0.153	**0.095**	**0.942**	219	**0.245**	0.084	0.839	296	0.239	0.084	0.845
20NewsGroups	294	**0.185**	0.098	0.913	267	0.178	**0.112**	**0.934**	259	0.160	0.092	0.932
Enron	224	**0.455**	**0.353**	0.898	298	0.315	0.346	**0.969**	271	0.255	0.292	0.963
NYTimes	125	**0.345**	**0.273**	**0.928**	173	0.319	0.216	0.897	196	0.292	0.199	0.907

is unique. However, in case of k-means and NG algorithms after obtaining the optimal k, we need to re-perform 10 randomizations as these techniques start with random initialization. We again use the entire text corpus as test data for these algorithms and perform these randomizations for the optimal k. In our experiments, we report the mean cluster quality values for k-means and NG algorithms. All our experiments were conducted on a desktop computer with 8Gb Ram, 2.4Ghz i7 intel processor using MATLAB 2013b.

B. Experimental Results

Table III showcases the results of proposed approach in comparison with traditional k-means and NG algorithms for several real world text datasets. From Table III, we can observe a trend that the KSDC method favours the Micro F-measure (F_m) i.e. KSDC method tends to generate more number of smaller sized homogeneous clusters as can be seen for Nips, Kos, Classic4, TDT2, TDT2all, Enron and NYTimes datasets. We can also observe that k-means is biased towards Macro F-measure (F_M) and generates fewer number of relatively large sized clusters as observed from Nips, Kos, R8, Reuters, TDT2 and TDT2all datasets. The k-means approach emerges as the prime competitor to the KSDC methodology as shown in Table III and the NG approach usually performs the worst.

From Table III, it can also be depicted that the KSDC method optimizes Q metric better than k-means and NG algorithms. This follows from the KSDC model as the Q metric is used as a criterion to locate the optimal k during the validation stage. Thus, KSDC model generally finds a good balance between the homogeneity of the clusters and preventing large-sized heterogeneous clusters of documents. We define an additional criterion namely balance which measures the average size of all the clusters obtained by a particular algorithm with the maximum sized cluster obtained by that method such that:

$$bal = \frac{\frac{\sum_{c_i \in \mathcal{C}} |c_i|}{|\mathcal{C}|}}{\text{argmax}_{c_i \in \mathcal{C}} |c_i|}. \tag{12}$$

This criterion can take values between $[0, 1]$ where higher values indicate more uniform sized clusters. We measure this criterion for the results obtained by KSDC, k-means and NG algorithm for these text datasets in Table IV. From Table IV, we can observe that the KSDC methodology mostly generates more balanced clusters in comparison to k-means and NG algorithms. This follows from the Q metric which tries to balance the homogeneity of the clusters (F_M) with the prevention of heterogeneous degenerate clusters (F_m).

TABLE IV: Comparison of KSDC, k-means and NG algorithms w.r.t. the balance (bal) of the clusters obtained for 10 real world text datasets.

Dataset	KSDC (bal)	k-means (bal)	NG (bal)
Nips	**0.278**	0.212	0.243
Kos	**0.326**	0.244	0.275
R8	0.393	**0.423**	0.288
Classic4	**0.385**	0.298	0.322
Reuters	0.1667	0.194	**0.234**
TDT2	**0.375**	0.324	0.256
TDT2all	**0.355**	0.278	0.295
20NewsGroups	**0.411**	0.393	0.402
Enron	0.212	0.188	**0.225**
NYtimes	**0.342**	0.288	0.326

We performed an additional experiment to identify groups of documents which are clustered together by different clustering algorithms for the Nips dataset. From the bag-of-words which appear in these clustered documents we can estimate the homogeneity of the clusters. We provide the title of Nips papers for one of the document clusters obtained by KSDC, k-means and NG clustering methods in Table V.

TABLE V: Title of Nips papers belonging to one of the document clusters obtained by KSDC, k-means and Neural Gas algorithms respectively. KSDC and NG algorithms find a homogeneous cluster of 3 documents related to speech recognition. However, the k-means clustering method discovers a document cluster which is more heterogeneous as it also includes documents related to continuous speech recognition.

Algorithm	Document Clusters
KSDC	"Speech Recognition: Statistical and Neural Information Processing Approaches", "HMM Speech Recognition with Neural Net Discrimination", "Improved Hidden Markov Model Speech Recognition Using Radial Basis Function Networks".
k-means	"Speech Recognition: Statistical and Neural Information Processing Approaches", "A Continuous Speech Recognition System Embedding MLP into HMM", "Connectionist Approaches to the Use of Markov Models for Speech Recognition", "Continuous Speech Recognition by Linked Predictive Neural Networks", "Speech Recognition Using Connectionist Approaches", "Multi-State Time Delay Networks for Continuous Speech Recognition".
NG	"Speech Recognition: Statistical and Neural Information Processing Approaches", "Connectionist Approaches to the Use of Markov Models for Speech Recognition", "Connectionist Approaches to the Use of Markov Models for Speech Recognition".

V. Conclusion

In this paper, we utilized unsupervised quality metrics based on precision and recall [1] to come up with a new quality metric Q. This metric was used to optimize the model selection procedure of kernel spectral clustering (KSC) in order to perform document clustering (KSDC). We compared the quality of the resulting clusters obtained by KSDC with k-means and neural gas clustering algorithms. We observed that the KSDC method in general results in more balanced small-sized homogeneous document clusters (high Q) and prevents formation of degenerate large-sized heterogeneous clusters (high F_m).

Acknowledgment

This work was supported by EU: ERC AdG A-DATADRIVE-B (290923), Research Council KUL: GOA/10/-/09 MaNet , CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants-Flemish Government; FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants; IWT: projects: SBO POM (100031); PhD/Postdoc grants; iMinds Medical Information Technologies SBO 2014-Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017).

References

[1] J.C. Lamirel, P. Cuxac, R. Mall and G. Safi. "A New Efficient and Unbiased Approach for Clustering Quality Approach", *PAKDD 2011 Workshops*, pp. 209-220, 2012.

[2] D. Cutting, D. Karger, J. Pederson and J. Tukey. "Scatter-Gather: A Cluster-based Approach to Browsing Large Document Collections", *ACM SIGIR Conference*, 1992.

[3] L. Baker and A. McCallum. "Distributional Clustering of Words for Text Classification", *ACM SIGIR Conference*, 1998.

[4] R. Bekkerman, R. El-Yaniv, Y. Winter and N. Tishby. "On Feature Distributional Clustering for Text Categorization", *ACM SIGIR Conference*, 2001.

[5] C.C. Aggarwal and C.X. Zhai. *A Survey of Text Clustering Algorithms*, Mining Text Data, Springer, 2012.

[6] R. Xu and D. Wunsch. "Survey of Clustering Algorithms", *IEEE Transactions on Neural Networks and Learning Systems*, 16(3), pp. 645-678, 2015.

[7] A.Y. Ng, M.I. Jordan and Y. Weiss. "On spectral clustering: analysis and an algorithm", *In Proc. of the Advances in Neural Information Processing Systems*; Dietterich, T.G., Becker, S., Ghahramani, Z., editors, MIT Press: Cambridge, MA, pp. 849-856, 2002.

[8] U. von Luxburg. "A tutorial on Spectral clustering", *Statistical Computing*, Vol. 17, pp. 395-416, 2007.

[9] F.R.K. Chung. *Spectral Graph Theory*, American Mathematical Society, 1997.

[10] W. Xu, X. Liu and Y. Gong. "Document clustering based on non-negative matrix factorization", *ACM SIGIR Conference*, 2003.

[11] 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), pp. 335-347, 2010.

[12] R. Mall and J.A.K. Suykens. "Very Sparse LSSVM Reductions for Large-Scale Data", *IEEE TNNLS*, 10.1109/TNNLS.2014.2333879, 2014.

[13] R. Mall and J.A.K. Suykens. "Sparse Reductions for Fixed-Size Least Squares Support Vector Machines on Large Scale Data", *PAKDD*, pp. 161-173, 2013.

[14] R. Mall, R. Langone and J.A.K. Suykens. "Kernel Spectral Clustering for Big Data Networks", *Entropy, Special Issue: Big Data*, 15(5), pp. 1567-1586, 2013.

[15] R. Mall, R. Langone and J.A.K. Suykens. "FURS:Fast and Unique Representative Subset selection retaining large scale community structure", *Social Network Analysis and Mining*, 3(4), pp. 1075-1095, 2013.

[16] R. Mall, R. Langone and J.A.K. Suykens. "Self-Tuned Kernel Spectral Clustering for Large Scale Networks", *IEEE International Conference on Big Data (IEEE BigData)*, pp. 385-393, 2013.

[17] R. Mall, R. Langone and J.A.K. Suykens. "Multilevel Hierarchical Kernel Spectral Clustering for Real-Life Large Scale Complex Networks", *PLoS ONE*, 9(6):e99966, 2014.

[18] J.C. Lamirel, S. Al-Shehabi, C. Francois and M. Hofmann. "New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping", *Scientometrics*, Vol. 60, pp. 445-562, 2004.

[19] R. Rabbany, M. Takaffoli, J. Fagnan, O.R. Zaiane and R.J.G.B. Campello. "Relative Validity Criteria for Community Mining Algorithms", *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 258-265, 2012.

[20] D. Forest. *Application de techniques de forage de textes de nature predictive et exploratoire 'a des fins de gestion et danalyse thematique de documents textuels non structur's*, PhD Thesis, Quebec University, Montreal, Canada, 2007.

[21] R. Kassab and J.C. Lamirel. "Feature Based Cluster Validation for High Dimensional Data", *In: IASTED International Conference on Artificial Intelligence and Applications (AIA)*, pp. 97-103, 2008.

[22] E. W. Forgy. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications", *Biometrics*, vol. 21, pp. 768-769, 1965.

[23] A. K. Jain. "Data Clustering: 50 years beyond k-means", *PR Letters*, 31(8), pp. 651-666, 2010.

[24] T. Martinetz and K. Schulten. *Artificial Neural Networks*, Elsevier, pp. 397-402, 1991.

[24] C.L. Blake and C.J. Merz. *UCI repository of machine learning databases*, http://archive.ics.uci.edu/ml/datasets.html, Irvine, CA, 1998.

[25] G. Salton and C. Buckley. "Term Weighting Approaches in Automatic Text Retrieval", *Information Processing and Management*, 24(5), pp. 513-523, 1988.

[26] L. Muflikhah. "Document Clustering using Concept Space and Cosine Similarity Measurement", *ICCTD*, pp. 58-62, 2009.

[27] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle. *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.

[28] R. Mall, V. Jumutcs, R. Langone and J.A.K. Suykens. "Representative subsets for big data learning using k-NN graphs", *IEEE BigData*, pp. 37-42, 2014.

[29] R. Langone, R. Mall and J.A.K. Suykens. "Soft Kernel Spectral Clustering", *IJCNN*, pp. 1-8, 2013.