Optimized data fusion for K-means Laplacian Clustering
Shi Yu, Xinhai Liu, León-Charles Tranchevent, Wolfgang Glänzel, Johan A. K. Suykens, Bart De Moor and Yves Moreau

1Institute for Genomics and Systems Biology, the University of Chicago.
2Department of Electrical Engineering, SISTA-SCD, Katholieke Universiteit Leuven.
3Steunpunt O&O Indicatoren, Katholieke Universiteit Leuven.

Prof. John Quackenbush

ABSTRACT
Motivation: We propose a novel algorithm to combine multiple kernels and Laplacians for clustering analysis. The new algorithm is formulated on a Rayleigh quotient objective function and is solved as a bi-level alternating minimization procedure. Using the proposed algorithm, the coefficients of kernels and Laplacians can be optimized automatically.
Results: Three variants of the algorithm are proposed. The performance is systematically validated on two real-life data fusion applications. The proposed OKLC algorithms perform significantly better than other methods. Moreover, the coefficients of kernels and Laplacians optimized by OKLC show some correlation with the rank of performance of individual data source. Though in our evaluation the K values are predefined, in practical studies, the optimal cluster number can be consistently estimated from the eigenspectrum of the combined kernel Laplacian matrix.
Availability: The MATLAB code of algorithms implemented in this paper is downloadable from http://homes.esat.kuleuven.be/~syu/oklc/oklc.html.
Contact: shee.yu@gmail.com

1 INTRODUCTION
Clustering is a fundamental problem in unsupervised learning and a number of different algorithms and methods have emerged over the years. K-means and spectral clustering are two popular methods for clustering analysis. K-means (KM) is proposed to cluster attribute-based data into K numbers of clusters with the minimal distortion (Bishop (2006); Duda et al. (2001)). Another well known method, spectral clustering (SC) (Ng et al. (2001); Shi and Malik (2000)), is also widely adopted in many applications. Unlike KM, SC is specifically developed for graphs, where the data samples are represented as vertices connected by non-negatively weighted undirected edges. The problem of clustering on graphs belongs to another paradigm than the algorithms based on the distortion measure. The goal of graph clustering is to find partitions on the graph such that the edges between different groups have a very low weight (von Luxburg (2007)). To model this, different objective functions are adopted and the typical criteria include the RatioCut (Hagen and Kahng (1992)), the normalized cut (Shi and Malik (2000)), and many others. To solve these objectives, the discrete constraint of the clustering indicators is usually relaxed to real values thus the approximated solution of spectral clustering can be obtained from the eigenspectrum of the graph Laplacian matrix. Many investigations (Dhillon et al. (2004)) have shown the connection between KM and SC. Moreover, in practical applications, the weighted similarity matrix is often used interchangeably as the kernel matrix in KM or the adjacency matrix in SC.

Recently, a new algorithm, Kernel Laplacian clustering (KL), is proposed to combine a kernel and a Laplacian simultaneously in clustering analysis (Wang et al. (2009)). This method combines the objectives of KM and SC in a quotient trace maximization form and solves the problem by eigen-decomposition. KL is shown to empirically outperforming KM and SC on real data sets. This straightforward idea is useful to solve many practical problems, especially those pertaining to combine attribute-based data with interaction-based networks. For example, in web analysis and scientometrics, the combination of text mining and bibliometrics has become a standard approach in clustering science or technology fields towards the detection of emerging fields or hot topics (Liu et al. (2010)). In bioinformatics, protein-protein interaction network and expression data are two of the most important sources used to reveal the relevance of genes and proteins with complex diseases. Conventionally, the data is often transformed into similarity matrices or interaction graphs, then consequently clustered by KM or SC. In KL, the similarity based kernel matrix and the interaction based Laplacian matrix are combined, which provides a novel approach to combine heterogeneous data structures in clustering analysis.

Our preliminary experiments show that when using KL to combine a single kernel and a single Laplacian, its performance strongly depends on the quality of the kernel and the Laplacian, which results in a model selection problem to determine the optimal settings of the kernel and the Laplacian. To perform model selection on unlabeled data is non-trivial because it is difficult to evaluate the models. To tackle the new problem, we propose a novel algorithm to incorporate multiple kernels and Laplacians in KL clustering. Our recent work proposes a method to integrate multiple kernel matrices in kernel k-means clustering (Yu et al. “Optimized data fusion for kernel K-means clustering”, submitted). The main contribution of the present work lies in the additive combination of multiple kernels and Laplacians, moreover, the coefficients assigned to the...
kernels and the Laplacians are optimized automatically. This paper presents the mathematical derivations of the additive integration form of kernels and Laplacians. The optimization of coefficients and clustering are achieved via a solution based on bi-level alternating minimization (Csiszar and Tusnady (1984)). We validate the proposed algorithm on heterogeneous data sets taken from two real applications, where the advantage and reliability of the proposed method are systematically compared and demonstrated.

2 METHODS

2.1 Combine kernel and Laplacian as generalized Rayleigh quotient for clustering

We first briefly review the KL algorithm proposed by (Wang et al. (2009)). All the mathematical symbols used in the paper are consistent and their representations are listed in Supplementary Material 1. Let us denote \( X \) as an attribute data set and \( W \) as a graph affinity matrix, both of them are representations of the same sets of samples. The objective of the KL integration to combine \( X \) and \( W \) for clustering can be defined as

\[
J_{KL} = \kappa J_{SC} + (1 - \kappa) J_{KM},
\]

where \( J_{SC} \) and \( J_{KM} \) are respectively the objectives of SC and KM clustering, \( \kappa \in [0, 1] \) is a coefficient adjusting the effect of the two objectives. Let us denote \( A \in \mathbb{R}^{N \times K} \) as the weighted scalar cluster membership matrix, given by

\[
A_{ab} = \begin{cases} \frac{1}{\sqrt{n_b}} & \text{if } \vec{x}_a \in C_b \\ 0 & \text{if } \vec{x}_a \notin C_b, \end{cases}
\]

where \( n_b \) is the number of data points belonging to cluster \( C_b \) and \( A^T A = I_K \), where \( I_K \) denotes a \( K \times K \) identity matrix. Let’s denote \( D \) as the diagonal matrix whose \((a, a)\) entry is the sum of the entries of row \( a \) in the affinity matrix \( W \). The normalized Laplacian matrix (von Luxburg (2007)) is given by

\[
\hat{L} = I - D^{-1/2} W D^{-1/2}.
\]

The objective of normalized cut based SC is formulated as

\[
\min_A \text{trace} \left( A^T \hat{L} A \right).
\]

As discussed in the literature (Duda et al. (2001); Bishop (2006); Hastie et al. (2009)), if the data \( X \) has zero sample means, the objective of the KM is given by

\[
\max_A \text{trace} \left( A^T X^T X A \right).
\]

We further generalize (5) by applying the feature map \( \phi(\cdot) : \mathbb{R} \to F \) on \( X \), then the centered data in \( F \) is denoted as \( X^\Phi \), given by

\[
X^\Phi = [\phi(\vec{x}_1) - \hat{\mu}^\Phi, \phi(\vec{x}_2) - \hat{\mu}^\Phi, ..., \phi(\vec{x}_N) - \hat{\mu}^\Phi],
\]

where \( \phi(\vec{x}_a) \) is the feature map applied on the column vector of the \( a \)-th data point in \( F \), \( \hat{\mu}^\Phi \) is the global mean in \( F \). The inner product \( X^\Phi X^\Phi^T \) in (5) can be combined using the kernel trick \( G(\vec{x}_a, \vec{x}_b) = \phi(\vec{x}_a) \phi(\vec{x}_b) \), where \( G(\cdot, \cdot) \) is a Mercer kernel. We denote \( G_c \) as the centered kernel matrix as \( G_c = PGP \), where \( P \) is the centering matrix \( P = I_N - (1/N)1_N1_N^T \), \( G \) is the kernel matrix, \( I_N \) is the \( N \times N \) identity matrix, \( 1_N \) is a column vector of \( N \) ones. Without loss of generality, the K-means objective in (5) can be equivalently written as

\[
\max_A \text{trace} \left( A^T G_c A \right).
\]

Then the objective of KL integration becomes

\[
\begin{align*}
\min_A & \quad \text{trace} \left( A^T \hat{L} A \right) - (1 - \kappa) \text{trace} \left( A^T G_c A \right) \\
\text{subject to} & \quad A^T A = I_K, \\
& \quad 0 \leq \kappa \leq 1.
\end{align*}
\]

To solve the optimization problem without tuning the ad hoc parameter \( \kappa \), Wang et al. formulate it as a trace quotient of the twocomponents (Wang et al. (2009)). The trace quotient is then further relaxed as a maximization of quotient trace, given by

\[
\max_A \text{trace} \left( A^T \hat{L} A \right)^{-1} (A^T G_c A) \\
\text{subject to } A^T A = I_K.
\]

The problem in (9) is a generalized Rayleigh quotient and the optimal solution \( A^* \) is obtained in the generalized eigenvalue problem. To maximize this objective, \( A^* \) is approximated as the largest \( K \) eigenvectors of \( \hat{L}^+ G_c \), where \( \hat{L}^+ \) is the pseudo inverse of \( \hat{L} \) (Wang et al. (2009)).

2.2 Combine kernel and Laplacian as additive models for clustering

As discussed, the original KL algorithm is proposed to optimize the generalized Rayleigh quotient objective. In this paper, we propose an alternative integration method using a different notation of Laplacian (von Luxburg (2007)), \( \tilde{L} \), given by

\[
\tilde{L} = D^{-1/2} W D^{-1/2},
\]

where \( D \) and \( W \) are defined the same as in (3). The objective of spectral clustering is equivalent to maximizing the term as

\[
\max_A \text{trace}(A^T \tilde{L} A).
\]

Therefore, the objective of the KL integration can be rewritten in an additive form, given by

\[
\begin{align*}
\min_A & \quad \kappa \text{trace} \left( A^T \tilde{L} A \right) + (1 - \kappa) \text{trace} \left( A^T G_c A \right) \\
\text{subject to} & \quad A^T A = I_K, \\
& \quad 0 \leq \kappa \leq 1,
\end{align*}
\]

where \( A, G_c \) are defined the same as in (8), \( \kappa \) is the free parameter to adjust the effect of kernel and Laplacian in KL integration. If \( \kappa \) is pre-defined, (12) is a Rayleigh quotient problem and the optimal \( A^* \) can be obtained from eigenvalue decomposition, known as the spectral relaxation (Ding and He (2004)). Therefore, to maximize this objective, we denote \( \Omega = \kappa \tilde{L} + (1 - \kappa) G_c \) thus \( A^* \) is solved as the dominant \( K \) eigenvectors of \( \Omega \).

In section 2.1 and section 2.2, two different methods are investigated to integrate a single Laplacian matrix with a single kernel matrix for clustering, where the main difference is to either optimize the cluster assignment affinity matrix \( A \) as a generalized Rayleigh quotient (ratio model) or as a Rayleigh quotient (additive model). The main advantage of the ratio based solution is to avoid tuning the parameter \( \kappa \). However, since the main contribution of this paper is to optimize the combination of multiple kernels and Laplacians, the coefficients assigned on each kernel and Laplacian still need to be optimized. Moreover, the optimization of the additive integration model is computationally simpler than optimizing the ratio based model. Therefore, in the following sections we will focus on extending the additive KL integration to multiple sources.

2.3 Clustering by multiple kernels and Laplacians: an additive model solved with bi-level optimization

Let us denote a set of graphs as \( H_i, i \in \{1, ..., r\} \), all having \( N \) vertices, and a set of Laplacians \( L_i \) constructed from \( H_i \) as in (10). Let us also denote a set of centered kernel matrices as \( G_{cj}, j \in \{1, ..., s\} \) with \( N \)
samples. To extend (12) by incorporating multiple kernels and Laplacians for clustering, we propose a strategy to learn their optimal weighted convex linear combinations. The extended objective function is then given by

$$Q1: \begin{array}{l}
\text{maximize } J_{Q1} = \text{trace } (A^T (\hat{L} + G) A) \\
\text{subject to } \hat{L} = \sum_{i=1}^{r} \theta_i \hat{L}_i, \\
G = \sum_{j=1}^{s} \theta_{j+r} G_{cj}, \\
\sum_{i=1}^{r} \theta_i^2 = 1, \quad \sum_{j=1}^{s} \theta_{j+r}^2 = 1, \\
\theta_i \geq 0, \quad l = 1, \ldots, (r + s), \\
A^T A = I_K,
\end{array}$$

where $\theta_1, \ldots, \theta_r$ and $\theta_{r+1}, \ldots, \theta_{r+s}$ are respectively the optimal coefficients assigned to the Laplacians and the kernels. $G$ and $L$ are respectively the combined kernel matrix and the combined Laplacian matrix. The $\kappa$ parameter in (12) is replaced by the coefficients assigned on each individual data sources.

To solve $Q1$, in the first phase we maximize $J_{Q1}$ with respect to $A$, keeping $\hat{\theta}$ fixed (initialized by random guess). In the second phase we maximize $J_{Q1}$ with respect to $\hat{\theta}$, keeping $A$ fixed. The two-stage optimization is then repeated until convergence. When $\hat{\theta}$ is fixed, denoting $\Omega = \hat{L} + G$, $Q1$ is exactly a Rayleigh quotient problem and the optimal $A^*$ can be solved as a eigenvalue problem of $\Omega$. When $A$ is fixed, the problem reduces to the optimization of the coefficients $\hat{\theta}$ with given cluster memberships. In Supplementary material 2 we prove that when the $A$ is given, $Q1$ can be formulated as Kernel Fisher Discriminant (KFD) in the high dimensional feature space $F$. We introduce $W = [\bar{w}_1, \ldots, \bar{w}_K]$, a projection matrix determining the pairwise discriminating hyperplane. Since the discriminant analysis is invariant to the magnitude of $\bar{w}_i$, we assume that $W^T W = I_K$, thus $Q1$ can be equivalently formulated as

$$Q2: \begin{array}{l}
\text{maximize } J_{Q2} = \text{trace } (W^T A^T A W)^{-1} (W^T A^T (G + \hat{L}) A W) \\
\text{s.t. } A^T A = I_K, \\
W^T W = I_K, \\
\hat{L} = \sum_{i=1}^{r} \theta_i \hat{L}_i, \\
G = \sum_{j=1}^{s} \theta_{j+r} G_{cj}, \\
\theta_i \geq 0, \quad l = 1, \ldots, (r + s), \\
\sum_{i=1}^{r} \theta_i^2 = 1, \quad \sum_{j=1}^{s} \theta_{j+r}^2 = 1.
\end{array}$$

The bi-level optimization to solve $Q1$ corresponds to two steps to solve $Q2$. In the first step (clustering), we set $W = \hat{L}_k$ and optimize $A$, which is exactly the additive kernel Laplacian integration as (12); In the second step (KFD), we fix $A$ and optimize $W$ and $\hat{\theta}$. Therefore, the two components optimize towards the same objective as a Rayleigh quotient in $F$ so the iterative optimization converges to a local optimum. Moreover, in the second step, we are not interested in the separating hyperplane defined in $W$, instead, we only need the optimal coefficients $\hat{\theta}$ assigned on the Laplacians and the kernels. It is known that Fisher Discriminant Analysis is related to the least squares approach (Duda et al. (2001)), and the KFD is related to and can be solved as the least squares formulation of support vector machines (Mika et al. (1999)), known as least squares Support Vector Machines (LSSVM) proposed by (Suykens et al. (2002)). The problem of optimizing multiple kernels for supervised learning (MKL) has been studied by (Lanckriet et al. (2004); Bach et al. (2004)). In our recent work (Yu et al. (2010b)), we derive the MKL extension for LSSVM and propose some efficient solutions to solve the problem. In this paper, the KFD problems are formulated as LSSVM MKL and solved by Semi-infinite programming (SIP) (Sonnenburg et al. (2006)). The concrete solutions and algorithms are presented in (Yu et al. (2010b)).

2.3.1 Optimize $A$ with given $\theta$ When $\theta$ are given, the kernel-Laplacian combined matrix $\Omega$ is also fixed, therefore, the optimal $A$ can be found as the dominant $K$ number of eigenvectors of $\Omega$.

2.3.2 Optimize $\theta$ with given $A$ When $A$ is given, the optimal $\theta$ assigned on Laplacians can be solved via the following KFD problem

$$Q3: \begin{array}{l}
\text{maximize } J_{Q3} = \text{trace } (W^T A^T A W)^{-1} (W^T A^T \hat{L} A W) \\
\text{s.t. } W^T W = I_K, \\
\hat{L} = \sum_{i=1}^{r} \theta_i \hat{L}_i, \\
\theta_i \geq 0, \quad i = 1, \ldots, r, \\
\sum_{i=1}^{r} \theta_i^2 = 1.
\end{array}$$

In our recent work, we have found that the $\delta$ parameter controls the sparseness of source coefficients $\theta_1, \ldots, \theta_r$ (Yu et al. (2010b)). When $\delta$ is set to 1, the optimized solution is sparse, which assigns dominant values to only one or two Laplacians (kernels) and zero values to the others. The sparseness is useful to distinguish relevant sources from a large number of irrelevant data sources. However, in many applications, there are usually a small number of sources and most of these data sources are carefully selected and preprocessed. They thus often are directly relevant to the problem. In these cases, a sparse solution may be too selective to thoroughly combine the complementary information in the data sources. We may thus expect a non-sparse integration method which smoothly distributes the coefficients on multiple kernels and Laplacians and, at the same time, leverages their effects in the objective optimization. We have proved that when $\delta$ is set to 2, the KFD step in (15) optimizes the $L_2$-norm of multiple kernels, which yields a non-sparse solution. If we set $\delta$ to 0, the cluster objective is simplified as to averagely combine multiple kernels and Laplacians. In this paper, we set $\delta$ to three different values ($0, 1, 2$) to respectively optimize the sparse, average, and non-sparse coefficients on kernels and Laplacians.

When $\delta$ is set to 1, the KFD problem in $Q3$ is solved as LSSVM MKL (Yu et al. (2010b)), given by

$$Q4: \begin{array}{l}
\text{minimize } J_{Q4} = \frac{1}{2} \sum_{i}^{r} \frac{1}{2 \lambda} \sum_{b=1}^{K} \gamma_{b}^2 \beta_{b}^2 - \sum_{b=1}^{K} \beta_{b}^2 Y_b^{\text{diag}}^{-1} \beta_{b} \\
\text{s.t. } \sum_{a=1}^{N} \beta_{ab} = 0, \quad b = 1, \ldots, K, \\
t \geq \sum_{b=1}^{K} \beta_{b}^2 L_b \beta_{b}, \quad i = 1, \ldots, r, \quad b = 1, \ldots, K,
\end{array}$$

where $\beta$ is the vector of dual variables, $t$ is a dummy variable in optimization, $a$ is the index of data samples, $b$ is the cluster label index of the discriminating problem in KFD, $Y_b$ is the diagonal matrix representing the
binary cluster assignment, the vector on the diagonal of $Y_i$ is equivalent to the $b$-th column of a affinity matrix $F_{ab}$ using $\{+1, -1\}$ to discriminate the cluster assignments, given by

$$F_{ab} = \begin{cases} +1 & \text{if } A_{ab} > 0, a = 1, \ldots, N, b = 1, \ldots, K \\ -1 & \text{if } A_{ab} = 0, a = 1, \ldots, N, b = 1, \ldots, K \end{cases} \quad (17)$$

The problem presented in Q4 has an efficient solution based on Semi-infinite programming (SIP), which is presented in equation forty-one of (Yu et al. 2010b)). The optimal coefficients $\theta_i$ correspond to the dual variables bounded by the quadratic constraint $\sum_{b=1}^{K} \beta_b^2 Y_b Y_b^T \leq 1$. The problem Q5 also has an efficient solution presented in equation forty-two in our recent work (Yu et al. 2010a)). The data sets contain nine different gene-by-term text profiles indexed by nine controlled vocabularies.

The problem presented in Q4 has an efficient solution based on Semi-infinite programming (SIP), which is presented in equation forty-one of (Yu et al. 2010b)). The optimal coefficients $\theta_i$ correspond to the dual variables bounded by the quadratic constraint $\sum_{b=1}^{K} \beta_b^2 Y_b Y_b^T \leq 1$. The problem Q5 also has an efficient solution presented in equation forty-two in our recent work (Yu et al. 2010a)). The data sets contain nine different gene-by-term text profiles indexed by nine controlled vocabularies.

**Algorithm 2.1: OKLC($G_{c1}, \ldots, G_{cs}, L_{1}, \ldots, L_{r}, K$)**

**comment:** Obtain the $\Omega^{(0)}$ using the initial guess of $\theta_1^{(0)}, \ldots, \theta_r^{(0)}$

$$A^{(0)} \leftarrow \text{EIGENVALUE DECOMPOSITION}(\Omega^{(0)}, K)$$

$\gamma = 0$

while $(\Delta A > \epsilon)$

1. $F^{(\gamma)} = F^{(\gamma-1)}$
2. $\theta_1^{(\gamma)}, \ldots, \theta_r^{(\gamma)} \leftarrow \text{SIP-LSSVM-MKL}(L_{1}, \ldots, L_{r}, F^{(\gamma)})$
3. $\theta_1^{(\gamma)}, \ldots, \theta_r^{(\gamma)} \leftarrow \text{SIP-LSSVM-MKL}(G_{c1}, \ldots, G_{cs}, F^{(\gamma)})$
4. $L^{(\gamma+1)} = L^{(\gamma)}$
5. $A^{(\gamma+1)} \leftarrow \text{EIGENVALUE DECOMPOSITION}(\Omega^{(\gamma+1)}, K)$
6. $\Delta A = \|A^{(\gamma+1)} - A^{(\gamma)}\|^2 / \|A^{(\gamma)}\|^2$
7. $\gamma := \gamma + 1$

return $(A^{(\gamma)}, \theta_1^{(\gamma)}, \ldots, \theta_r^{(\gamma)}, \theta_1^{(\gamma+1)}, \ldots, \theta_r^{(\gamma+1)})$

**Algorithm 2.1:** OKLC($G_{c1}, \ldots, G_{cs}, L_{1}, \ldots, L_{r}, K$)

**comment:** Obtain the $\Omega^{(0)}$ using the initial guess of $\theta_1^{(0)}, \ldots, \theta_r^{(0)}$

$A^{(0)} \leftarrow \text{EIGENVALUE DECOMPOSITION}(\Omega^{(0)}, K)$$\gamma = 0$

while $(\Delta A > \epsilon)$

1. $F^{(\gamma)} = F^{(\gamma-1)}$
2. $\theta_1^{(\gamma)}, \ldots, \theta_r^{(\gamma)} \leftarrow \text{SIP-LSSVM-MKL}(L_{1}, \ldots, L_{r}, F^{(\gamma)})$
3. $\theta_1^{(\gamma)}, \ldots, \theta_r^{(\gamma)} \leftarrow \text{SIP-LSSVM-MKL}(G_{c1}, \ldots, G_{cs}, F^{(\gamma)})$
4. $L^{(\gamma+1)} = L^{(\gamma)}$
5. $A^{(\gamma+1)} \leftarrow \text{EIGENVALUE DECOMPOSITION}(\Omega^{(\gamma+1)}, K)$
6. $\Delta A = \|A^{(\gamma+1)} - A^{(\gamma)}\|^2 / \|A^{(\gamma)}\|^2$
7. $\gamma := \gamma + 1$

return $(A^{(\gamma)}, \theta_1^{(\gamma)}, \ldots, \theta_r^{(\gamma)}, \theta_1^{(\gamma+1)}, \ldots, \theta_r^{(\gamma+1)})$

**2.4 Data sets and experimental setup**

The proposed OKLC models are validated in two real applications to combine heterogeneous data sets in clustering analysis. The data sets in the first experiment is taken from the work of multi-view text mining for disease gene identification (Yu et al. 2010a)). The data sets contain nine different gene-by-term text profiles indexed by nine controlled vocabularies.

The original disease relevant gene data set contains 620 genes which are known to be relevant to 29 diseases. To avoid the effect of imbalanced clusters which may affect the evaluation, we only keep the diseases who have 11 to 40 relevant genes. This results in 14 genetic diseases and 278 genes. Because the present paper is focused on non-overlapping (“hard”) clustering, we further remove 16 genes which are relevant to multiple diseases.

The remaining 262 disease relevant genes are clustered into 14 clusters and evaluated biologically by their disease labels. For each vocabulary based gene-by-term data source, we create a kernel matrix using the linear kernel function and the kernel normalization method proposed by (Shawe-Taylor and Cristianini 2004)). An element in the kernel matrix is then equivalent to the value of cosine similarity of two vectors (Baeza-Yates and Ribeiro-Neto 1999)). This kernel is then regarded as the weighted adjacency matrix to create the Laplacian matrix. In total, nine kernels and nine Laplacian matrices are combined in clustering.

The data sets in the second experiment are taken from Web of Science (WOS) database provided by Thomson Scientific (Liu et al. 2010)). After preprocessing, the data set contains 8305 journals categorized in 22 scientific fields. To create a balanced benchmark data for evaluation, we select seven fields consisting 1421 journals. The titles, abstracts and keywords of the journal publications are indexed by a text mining program using no controlled vocabulary. The weights of terms are calculated using four weighting schemes: TF-IDF, IDF, TF and binary. The citations among journals are also investigated from four different aspects: cross-citation, co-citation, bibliographic coupling and binary cross-citation. The lexical similarities are represented as normalized linear kernel matrices (using the same methods applied on the disease data) and the citation metrics are regarded as weighted adjacency matrices to create the Laplacians. Totally four kernels and four Laplacians are combined on journal data. The details about the two data sets are presented in Supplementary material 3.

The data sets used in our experiments are provided with labels, therefore the clustering performance is evaluated as comparing the automatic partitions with the labels using Adjusted Rand Index (ARI) (Hubert and Arabie 1985)) and Normalized Mutual Information (NMI) (Strehl and Ghosh 2002)). To evaluate the ARI and NMI performance, we set $K = 14$ on disease data and $K = 7$ on journal data. We also tune the OKLC model using different K values.
To evaluate whether the combination of kernel and Laplacian indeed improves the clustering performance, we first systematically compare the performance of all the individual data source using KM and SC. As shown in Supplementary material 4, on disease data, the best KM performance (ARI 0.5441, NMI 0.7099) and SC (ARI 0.5199, NMI 0.6858) performance is obtained on LDDDB text mining profile. Next, we enumerate all the paired combinations of a single kernel and a single Laplacian for clustering. The integration is based on equation (12) and the $\kappa$ value is set to 0.5 so the objectives of KM and SC are combined averagely.

### 3 RESULTS

We implement the proposed OKLC models to integrate multiple kernels and Laplacians on disease data and journal set data. To compare the performance, we also apply six popular ensemble clustering methods mentioned in relevant work (Yu et al., 2010a)) to combine the partitions of individual kernels and Laplacians as a consolidate partition. These six methods are CSPA (Strehl and Ghosh (2002)), HGPA (Strehl and Ghosh (2002)), MCLA (Strehl and Ghosh (2002)), QMI (Topchy et al., 2005), EACAL (Fred and Jain (2005)), and AdacVote (Ayad and Kamel (2008)). As shown in Table 1 and Table 2, the performance of OKLC algorithms is better than all the compared methods and the improvement is significant. The performance of all 45 paired combinations is presented in Supplementary material 5. As shown, the best KL clustering performance is obtained by integrating the LDDDB kernel with KO Laplacian (ARI 0.5298, NMI 0.6949). Moreover, we also found that the integration performance varies significantly by the choice of kernel and Laplacian, which proves our previous point that the KL performance is highly dependent on the quality of kernel and Laplacian. Using the proposed OKLC algorithm, there is no need to enumerate all the possible paired combinations. OKLC combines all the kernels and Laplacians and optimizes their coefficients in parallel, yielding a comparable performance with the best paired combination of a single kernel and a single Laplacian.

In Figure 2 the confusion matrices of disease data for a single run are depicted. The values on the matrices are normalized according to $R_{ij} = C_i / T_i$, where $T_i$ is the total number of genes belonging in disease $i$ and $C_j$ is the number of these $T_i$ genes that were clustered to belong to class $j$. First, it is worth noting that OKLC reduce the number of mis-clustered genes on breast cancer (Nr.1), cardiomyopathy (Nr.2), and muscular dystrophy (Nr.11). Among the mis-clustered genes in LDDDB, five genes (TSG101, DBC1, CTTN, SLC22A18, AR) in breast cancer, two genes in cardiomyopathy (COX15, CRSP3), and two genes in muscular dystrophy (SEPN1, COL6A3) are correctly clustered in OKLC model 1. Second, there are several diseases where consistent misclustering occurs in both methods, such as diabetes (Nr.6) and neuropathy (Nr.12). The intuitive confusion matrices correspond to the numerical evaluation results, as shown, the quality of clustering obtained by OKLC model 1 (ARI=0.5898, NMI=0.7429) is higher than LDDDB.

The performance of individual data sources of journal data is shown in Supplementary material 6. The best KM (ARI 0.6482, NMI 0.7104) is obtained on the IDF kernel and the best SC (ARI 0.5667, NMI 0.6807) is obtained on the cross-citation Laplacian. To combine the 4 kernels with 4 Laplacians, we evaluate all the 10 paired combinations and show the performance in Supplementary material 7. The best performance is obtained by integrating the IDF kernel with the cross-citation Laplacian (ARI 0.7566, NMI 0.7702). As shown, the integration of lexical similarity information and citation based Laplacian indeed improves the performance.

In Figure 2 the confusion matrices (also normalized) of journal data for a single run are illustrated. We compare the best individual data source (IDF with kernel K-means, figure on the left) with the OKLC model 1. In the confusion matrix of IDF K-means, 79 journals belonging to agriculture science (Nr.1) are mis-clustered to environment ecology (Nr.3), 9 journals are mis-clustered to pharmacology and toxicology (Nr.7). In OKLC, the number of agriculture journals mis-clustered to environment ecology is reduced to 45, and the number to pharmacology and toxicology is reduced to 5. On other journal clusters, the performance of the two models is almost equivalent.

We also investigate the performance of only combining multiple kernels or multiple Laplacians. On the disease data set, we respectively combined the 9 kernels and the 9 Laplacians for

---

**Table 1.** Performance on disease data set. All the comparing methods combine nine kernels and nine Laplacians. The mean values and the standard deviations are observed from 20 random repetitions. The best performance is shown in bold. The P-values are statistically evaluated with the best performance using paired t-test.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>ARI</th>
<th>P-value</th>
<th>NMI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OKLC 1</td>
<td>0.5858</td>
<td>±0.0390</td>
<td>0.7451</td>
<td>±0.0194</td>
</tr>
<tr>
<td>OKLC 2</td>
<td>0.5369</td>
<td>±0.0493</td>
<td>0.7106</td>
<td>±0.0283</td>
</tr>
<tr>
<td>OKLC 3</td>
<td>0.5469</td>
<td>±0.0485</td>
<td>1.2300</td>
<td>±0.0360</td>
</tr>
<tr>
<td>CSFA</td>
<td>0.4367</td>
<td>±0.0266</td>
<td>5.6611</td>
<td>±0.0222</td>
</tr>
<tr>
<td>HGPA</td>
<td>0.5040</td>
<td>±0.0363</td>
<td>8.4700</td>
<td>±0.0307</td>
</tr>
<tr>
<td>MCLA</td>
<td>0.4731</td>
<td>±0.0320</td>
<td>2.2600</td>
<td>±0.0210</td>
</tr>
<tr>
<td>QMI</td>
<td>0.4656</td>
<td>±0.0425</td>
<td>7.7000</td>
<td>±0.0255</td>
</tr>
<tr>
<td>EACAL</td>
<td>0.4817</td>
<td>±0.0263</td>
<td>2.5000</td>
<td>±0.0109</td>
</tr>
<tr>
<td>AdacVote</td>
<td>0.1394</td>
<td>±0.0649</td>
<td>1.4700</td>
<td>±0.0740</td>
</tr>
</tbody>
</table>

**Table 2.** Performance on journal data set. All the comparing methods combine four kernels and four Laplacians. The mean values and the standard deviations are observed from 20 random repetitions. The best performance is shown in bold. The P-values are statistically evaluated with the best performance using paired t-test.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>ARI</th>
<th>P-value</th>
<th>NMI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OKLC 1</td>
<td>0.7346</td>
<td>±0.0584</td>
<td>0.3585</td>
<td>±0.0364</td>
</tr>
<tr>
<td>OKLC 2</td>
<td>0.7235</td>
<td>±0.0660</td>
<td>0.9732</td>
<td>±0.0358</td>
</tr>
<tr>
<td>OKLC 3</td>
<td>0.7336</td>
<td>±0.0499</td>
<td>0.7587</td>
<td>±0.0362</td>
</tr>
<tr>
<td>CSFA</td>
<td>0.6703</td>
<td>±0.0485</td>
<td>8.8405</td>
<td>±0.0291</td>
</tr>
<tr>
<td>HGPA</td>
<td>0.6763</td>
<td>±0.0419</td>
<td>4.7400</td>
<td>±0.0269</td>
</tr>
<tr>
<td>MCLA</td>
<td>0.6571</td>
<td>±0.0746</td>
<td>6.5500</td>
<td>±0.0493</td>
</tr>
<tr>
<td>QMI</td>
<td>0.6592</td>
<td>±0.0593</td>
<td>5.3200</td>
<td>±0.0326</td>
</tr>
<tr>
<td>EACAL</td>
<td>0.5808</td>
<td>±0.0178</td>
<td>3.8500</td>
<td>±0.0153</td>
</tr>
<tr>
<td>AdacVote</td>
<td>0.5899</td>
<td>±0.0556</td>
<td>1.0200</td>
<td>±0.0325</td>
</tr>
</tbody>
</table>
clustering, using all the compared methods in Table 1 and Table 2. On the journal data set, we combine the 4 text mining kernels and the 4 citation Laplacians. The proposed OKLC method is simplified as only optimizing coefficients on Laplacians (step 2 in Algorithm 2.1) or kernels (step 3). As shown in Supplementary material 8, the performance of OKLC is also comparable to the best performance obtained either by kernel combination or Laplacian combination. In particular, of all the methods we compared, the best performance is all obtained on OKLC models or its simplified forms.

In the proposed algorithm, when optimizing the coefficients of kernels and Laplacians as (16) and (18), the regularization parameter \( \lambda \) is estimated automatically and jointly as a coefficient of identity matrix (Yu et al. (2010b)). To investigate whether the automatically estimated \( \lambda \) leads to the optimal performance, we tried 21 pre-defined \( \lambda \) values uniformly sampled on the \( \log \) scale from \( 2^{-10}, \ 2^{-9}, \ldots, \ 2^{10} \). Figure 3 shows the ARI and NMI performance obtained by these \( \lambda \) values on disease data. Figure 4 presents the ARI and NMI on journal data. As shown, on both data sets, the automatically optimized \( \lambda \) value in OKLC model 1 (blue dashed lines) yields comparable performance as the best pre-defined \( \lambda \) values (blue solid lines). Moreover, the difference between sparse (blue lines) and non-sparse (red lines) optimized kernel Laplacian integration is more significant on disease data than on journal data. In OKLC, the user is free to choose the \( \lambda \) value estimated automatically or set it with a predefined value. Our experimental result shows the automatic estimation yields near-optimal performance.

It is interesting to observe that the average combination model (OKLC model 2) performs quite well on the journal data set but not on the disease data set. This is probably because most of the

**Fig. 1.** Confusion matrices of disease data obtained by kernel K-means on LDDB (figure on the left) and OKLC model 1 integration (figure on the right). The numbers of cluster labels are consistent with the numbers of diseases presented in Supplementary material 3. In each row of the confusion matrix, the diagonal element represents the fraction of correctly clustered genes and the off-diagonal non-zero element represents the fraction of mis-clustered genes.

**Fig. 2.** Confusion matrices of journal data obtained by kernel K-means on IDF (figure on the left) and OKLC model 1 integration (figure on the right). The numbers of cluster labels are consistent with the numbers of ESI journal categories presented in Supplementary material 3. In each row, the diagonal element represents the fraction of correctly clustered journals and the off-diagonal non-zero element represents the fraction of mis-clustered journals.
In this paper we propose a new clustering approach, OKLC, to integrate multiple data sources in journal data set. In disease data set some data sources are noisy, thus the integration of disease data sources is a non-trivial task. We expect that the other two OKLC models (model 1 and model 3) optimize the coefficients assigned on the kernels and the Laplacians to leverage multiple sources in integration, at the same time, to increase the robustness of the combined model on combining relevant and irrelevant data sources. To evaluate whether the optimized weights assigned on individual sources have correlation with the performance, we compare the rank of coefficients with the rank of performance from Table 3 to Table 6. As shown, the largest coefficients correctly indicate the best individual data sources. It is worth noting that in multiple kernel learning, the rank of coefficients may not be 100% correlated with the rank of individual performance. In our experiments, the MeSH kernel gets the 2nd largest weights though its performance in evaluation is low. In MKL, it is usual that the best individual kernel found by cross-validation may not lead to a large weight when used in combination (Ye et al. (2008)). Kernel fusion combines multiple sources at a refined granularity, where the “moderate” kernels containing weak and insignificant information could complement to other kernels to compose a “good” kernel containing strong and significant information. Though such complementary information cannot be incorporated when cross-validation is used to choose a single best kernel, these “moderate” kernels are still useful when combined with other kernels (Ye et al. (2008)). Based on the ranks presented in Table 5 and Table 6, we calculate the spearman correlations between the ranks of weights and the ranks of performance on both data sets. The correlations of disease kernels, disease Laplacians, journal kernels and journal Laplacians are respectively 0.5657, 0.6, 0.8, and 0.4. In some relevant work, the average spearman correlations are mostly around 0.4 (Lanckriet et al. (2004); Ye et al. (2008)). Therefore, the optimal weights obtained in our experiments are generally consistent with the rank of performance.

As a spectral clustering algorithm, the optimal cluster number of OKLC can be estimated by checking the plot of eigenvalues (von Luxburg (2007)). To demonstrate this, we investigate the dominant eigenvalues of the optimized combination of kernels and Laplacians. In Figure 3, we compare the difference of three OKLC models with the pre-defined \( K \) (set as equal to the number of class labels). In practical research, one can predict the optimal cluster number by checking the “elbow” of the eigenvalue plot. As shown in Figure 3, the “elbow” in disease data is quite obvious at the number of 14. In journal data, the “elbow” is more likely to range from 6 to 12. All the three OKLC models show a similar trend on the eigenvalue plot. Moreover, in Supplementary material 9 we also compare the eigenvalue curves using different \( K \) values as input. As shown, the eigenvalue plot is quite stable w.r.t. the different inputs of \( K \), which means the optimized kernel and Laplacian coefficients are quite independent with the \( K \) value. This advantage enables a reliable prediction about the optimal cluster number by integrating multiple data sources.

To investigate the computational time, we benchmark OKLC algorithms with other clustering methods on the two data sets. As shown in Table 7, when optimizing the coefficients, OKLC algorithm (model 1 and model 3) spends longer time than the other methods to optimize the coefficients on the Laplacians and the kernels. However, the proposed algorithm is still efficient. Considering the fact that the proposed algorithm yields much better performance and more enriched information (the ranking of the individual sources) than other methods, it is worth spending extra computational complexity on a promising algorithm.

### Table 3. The average values of coefficients of kernels and Laplacians in disease data set optimized by OKLC model 1. The sources assigned with 0 coefficient are not presented. The performance is ranked by the average values of ARI and NMI evaluated on each individual sources (Supplementary Material 3).

<table>
<thead>
<tr>
<th>Rank of ( \theta )</th>
<th>Source</th>
<th>( \theta ) value</th>
<th>Performance rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LDDDB kernel</td>
<td>0.6113</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>MESH kernel</td>
<td>0.3742</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Uniprot kernel</td>
<td>0.0095</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Omim kernel</td>
<td>0.0050</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 4. The average values of coefficients of kernels and Laplacians in journal data set optimized by OKLC model 1. The sources assigned with 0 coefficient are not presented. The performance is ranked by the average values of ARI and NMI evaluated on each individual sources (Supplementary Material 5).

<table>
<thead>
<tr>
<th>Rank of ( \theta )</th>
<th>Source</th>
<th>( \theta ) value</th>
<th>Performance rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LDDDB Laplacian</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 5. The average values of coefficients of kernels and Laplacians in disease data set optimized by OKLC model 3.

<table>
<thead>
<tr>
<th>Rank of ( \theta )</th>
<th>Source</th>
<th>( \theta ) value</th>
<th>Performance rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LDDDB kernel</td>
<td>0.4578</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>MESH kernel</td>
<td>0.3495</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>OMIM kernel</td>
<td>0.3376</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>SNOMED kernel</td>
<td>0.3309</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>MPO kernel</td>
<td>0.3178</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>GO kernel</td>
<td>0.3175</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>eVOC kernel</td>
<td>0.3180</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>Uniprot kernel</td>
<td>0.3089</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>KO kernel</td>
<td>0.2143</td>
<td>9</td>
</tr>
</tbody>
</table>

### Table 6. The average values of coefficients of kernels and Laplacians in journal data set optimized by OKLC model 3.

<table>
<thead>
<tr>
<th>Rank of ( \theta )</th>
<th>Source</th>
<th>( \theta ) value</th>
<th>Performance rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LDDDB Laplacian</td>
<td>0.6861</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>MESH Laplacian</td>
<td>0.2799</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>OMIM Laplacian</td>
<td>0.2680</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>GO Laplacian</td>
<td>0.2645</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>eVOC Laplacian</td>
<td>0.2615</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Uniprot Laplacian</td>
<td>0.2572</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>SNOMED Laplacian</td>
<td>0.2559</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>MPO Laplacian</td>
<td>0.2476</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>KO Laplacian</td>
<td>0.2163</td>
<td>9</td>
</tr>
</tbody>
</table>

### 4 CONCLUSION

In this paper we propose a new clustering approach, OKLC, to optimize the combination of multiple kernels and Laplacians in
Fig. 3. Performance comparison of the predefined $\lambda$ value and the automatically estimated $\lambda$ value in OKLC. The ARI performance (figure on the left) and NMI performance (figure on the right) is shown. The blue circle curves show the performance of OKLC model 1. The red square curves show the performance of OKLC model 2. The dashed straight lines are the performance obtained by the automatic estimated $\lambda$ values.

Fig. 4. Similar to Figure 3, this figure compares the predefined $\lambda$ value and automatic $\lambda$ estimation on journal data. The ARI performance (figure on the left) and NMI performance (figure on the right) is shown.

Table 6. The average values of coefficients of kernels and Laplacians in journal data set optimized by OKLC model 3.

<table>
<thead>
<tr>
<th>Rank of $\theta$</th>
<th>Source</th>
<th>$\theta$ value</th>
<th>Performance rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IDF kernel</td>
<td>0.5389</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Binary kernel</td>
<td>0.4520</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>TF kernel</td>
<td>0.2876</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>TF-IDF kernel</td>
<td>0.2376</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>Bibliographic Laplacian</td>
<td>0.7106</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Cocitation Laplacian</td>
<td>0.5134</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Crosscitation Laplacian</td>
<td>0.4450</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Binarycitation Laplacian</td>
<td>0.1819</td>
<td>3</td>
</tr>
</tbody>
</table>

The objective of OKLC is formulated as a Rayleigh quotient function and is solved iteratively as a bi-level optimization procedure. In the simplest interface, the proposed algorithm only requires one input parameter, the cluster number $K$, from the user. Moreover, depending on user’s expectation to select the most relevant sources or to evenly combine all sources, the sparseness of coefficient vector $\theta$ can be controlled via the parameter $\delta$. In our paper we propose three variants of the OKLC algorithm and validate them on two real applications. The performance of clustering is systematically compared with a variety of algorithms and different experimental settings. The proposed OKLC algorithms perform significantly better than other methods. Moreover, the coefficients of kernels and Laplacians optimized by OKLC show strong correlation with the rank of performance of individual data source. Though in our evaluation the $K$ values are predefined, in practical studies, the optimal cluster number can be consistently
The parameter $K$ is set as equivalent as the reference label numbers.

Table 7. Comparison of CPU time of all algorithms. The reported values are averaged from 20 repetitions. The CPU time is evaluated on Matlab v7.6.0 + Windows XP2 installed on a Laptop computer with Intel Core 2 Duo 2.26G Hz and 2G memory.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>disease data (seconds)</th>
<th>journal data (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OKLC model 1</td>
<td>42.39</td>
<td>1011.4</td>
</tr>
<tr>
<td>OKLC model 2</td>
<td>0.19</td>
<td>13.27</td>
</tr>
<tr>
<td>OKLC model 3</td>
<td>37.74</td>
<td>577.51</td>
</tr>
<tr>
<td>CSPA</td>
<td>9.49</td>
<td>177.22</td>
</tr>
<tr>
<td>HGPA</td>
<td>10.13</td>
<td>182.51</td>
</tr>
<tr>
<td>MCLA</td>
<td>9.95</td>
<td>320.93</td>
</tr>
<tr>
<td>QMI</td>
<td>9.36</td>
<td>186.25</td>
</tr>
<tr>
<td>EACAL</td>
<td>9.74</td>
<td>205.59</td>
</tr>
<tr>
<td>AdacVote</td>
<td>9.22</td>
<td>172.12</td>
</tr>
</tbody>
</table>

estimated from the eigenspectrum of the combined kernel Laplacian matrix.

The proposed OKLC algorithm demonstrates the advantage of combining and leveraging information from heterogeneous data structures and sources. It is potentially useful in bioinformatics and many other application areas, where there is a surge of interest to integrate similarity based information and interaction based relationships in statistical analysis and machine learning.

REFERENCES