Data Min Knowl Disc
DOI 10.1007/s10618-014-0384-8

Weakly supervised nonnegative matrix factorization for user-driven clustering

Jaegul Choo · Changhyun Lee · Chandan K. Reddy · Haesun Park

Received: 3 March 2013 / Accepted: 3 September 2014

Abstract Clustering high-dimensional data and making sense out of its result is a challenging problem. In this paper, we present a weakly supervised nonnegative matrix factorization (NMF) and its symmetric version that take into account various prior information via regularization in clustering applications. Unlike many other existing methods, the proposed weakly supervised NMF methods provide interpretable and flexible outputs by directly incorporating various forms of prior information. Furthermore, the proposed methods maintain a comparable computational complexity to the standard NMF under an alternating nonnegativity-constrained least squares framework. By using real-world data, we conduct quantitative analyses to compare our methods against other semi-supervised clustering methods. We also present the use cases where the proposed methods lead to semantically meaningful and accurate clustering results by properly utilizing user-driven prior information.

Responsible editor: Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, Filip Železný.

J. Choo (✉) · H. Park
Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: jaegul.choo@cc.gatech.edu

H. Park
e-mail: hpark@cc.gatech.edu

C. Lee
Google Inc., Mountain View, CA 94043, USA
e-mail: khakice@gmail.com

C. K. Reddy
Wayne State University, Detroit, MI 48202, USA
e-mail: reddy@cs.wayne.edu

Published online: 05 October 2014

Keywords Nonnegative matrix factorization · Semi-supervised clustering · User-driven clustering · Regularization

1 Overview

Clustering high-dimensional data is a complex, challenging problem in that the data do not often have distinct clusters and that the resulting clusters are not always semantically meaningful to end-users (Aggarwal and Reddy 2013). Among numerous approaches proposed to improve this problem, semi-supervised approaches, which impose additional information to the clustering processes, have been actively studied and applied to many clustering methods such as k-means.

This paper presents semi-supervised approaches for nonnegative matrix factorization (NMF) (Lee and Seung 1999; Kim and Park 2007) as well as its symmetric version (Kuang et al. 2012). NMF has shown great performances for clustering in various domains including text mining and computer vision (Xu et al. 2003; Shahnaz et al. 2006; Li et al. 2007; Kuang et al. 2012), but its semi-supervised approaches, which could further improve the superior performance of NMF, have not received enough attention.

We present flexible and effective semi-supervised NMF methods for clustering: weakly supervised NMF (WS-NMF) and symmetric NMF (WS-SymNMF). By 'weak supervision,' we intend to make the solution of our methods reflect the prior information given by users. In this process, our methods allow users to flexibly control how strongly to impose such prior information in the final solution. Representing this idea using the regularization terms that penalize the differences between the resulting output and the prior information, we present novel formulations and algorithms for WS-NMF and WS-SymNMF.

Our methods can flexibly accommodate diverse forms of prior information, as follows:

Partial versus entire coverage of data. A typical semi-supervised learning setting assumes that the cluster label information is available for partial data. On the other hand, cluster labels may be available for the entire data, say, obtained from a different source, and one may want to weakly impose such information in the final clustering result. A representative piece of work in this category is evolutionary clustering (Chakrabarti et al. 2006; Chi et al. 2009), which tries to maintain the temporal coherence of the clustering result for time-evolving data given the clustering result (of the entire set of data items) at a previous time step. Evolutionary clustering usually incorporates this idea in the form of regularization terms.

Hard- versus soft-clustering information. The prior information about a cluster label of an individual data item can be either a single label, i.e., a hard-clustering label, or a vector of cluster membership coefficients, i.e., a soft-clustering label (Xie and Beni 1991). The latter provides richer information than the former by specifying how strongly a data item is relevant to a particular cluster.

Data- versus cluster-level information. The prior information about a cluster may directly represent the cluster characteristics instead of the cluster labels of individual data items (Alqadah et al. 2012). For instance, one might want to obtain the cluster

strongly/weakly related to particular features (e.g., a topic cluster closely related to particular keywords of his/her choice in the case of document data).

To evaluate the proposed methods in various settings, we conduct both quantitative and qualitative experiments. For the former, we compare our methods with several existing methods in traditional semi-supervised learning applications. For the latter, we show interesting use cases for real-world data sets, where WS-NMF and WS-SymNMF lead to semantically meaningful and accurate clustering results by utilizing user-driven prior knowledge.

The main contributions of our work are summarized as follows:

- Novel formulations of WS-NMF and WS-SymNMF that can flexibly incorporate prior knowledge for user-driven clustering.
- Algorithm development for the proposed methods with a comparable computational complexity to that of the standard NMF algorithm.
- Quantitative experimental comparisons to demonstrate the superiority of the proposed methods in semi-supervised clustering applications.
- Usage scenarios using real-world data sets in which the clustering result is improved by incorporating users' prior knowledge.

The rest of this paper is organized as follows. Section 2 discusses related work, and Sect. 3 presents the formulations and the algorithms of WS-NMF and WS-SymNMF. Section 4 presents experimental results, and Sect. 5 concludes the paper.

2 Related work

NMF has been an active research topic in machine learning and data mining. Owing to its innate interpretability and good performance in practice, there have been significant research efforts towards improving the performance of NMF in various applications.

Cai et al. (2011) have proposed a regularization based on the manifold approximated by k-nearest neighbor graphs. Another graph-based regularization approach (Guan et al. 2011) has tried to promote part-based bases as well as maximize discrimination between pre-given classes. The sparsity for solution robustness has been taken into account in the form of hard constraints (Hoyer 2004) as well as regularization terms (Kim and Park 2007).

In addition, semi-supervised clustering has been actively studied recently (Bilenko et al. 2004; Basu et al. 2004, 2008). In the case of NMF, various forms of semi-supervised formulations and algorithms that incorporate prior knowledge have been proposed. Li et al. (2007) have presented an NMF-based formulation for consensus clustering as well as semi-supervised clustering based on pairwise clustering constraints. Multi-label learning has also been solved via a constrained NMF formulation by modeling the correlation or similarity between clusters (Liu et al. 2006).

More recently, other advanced semi-supervised NMF methods have been proposed. Chen et al. (2008) have taken into account pairwise clustering constraints and have incorporated such information by increasing or decreasing the weights of the corresponding components in the input similarity matrix. Chen et al. (2010) have also proposed a method based on the metric learning for minimizing the distances between

must-links while maximizing those between cannot-links, followed by the NMF clustering on the space obtained from such a metric learning step. Wang et al. (2009) have proposed another NMF-based formulation that incorporates the pairwise constraints in the resulting distances of data items while preserving the local neighborhood relationships. Liu et al. (2012) have solved an NMF problem so that the supervised points in the same cluster can have identical representations in the NMF output. Lee et al. (2010) have built a joint factorization formulation for both the original feature space and the label space linked via a common factor matrix. Then, they perform k-means clustering on this common factor matrix to obtain the final clustering result.

Considering the close connection between NMF and a popular topic modeling method, latent Dirichlet allocation (LDA) (Blei et al. 2003), there have been many studies leveraging the prior knowledge in the LDA formulation. The topic modeling based on Dirichlet-multinomial regression has shown a versatile capability in handling additional features associated with document data along with their textual information (Mimno and McCallum 2012). Similarly, Zeng et al. have proposed multi-relational topic modeling under the Markov random field framework, which can take into account multiple relationships between data items, e.g., both the externally given and the inherent relationships (Zeng et al. 2009).

Unlike these existing methods, our method has significant advantages in terms of *interpretability* and *flexibility*. As for *interpretability*, our method imposes the prior knowledge in a way that directly affects the final clustering output of NMF. On the contrary, many of the above-described methods indirectly apply the prior knowledge to the formulation, for instance, by manipulating the input matrix (or the distance) or by introducing another latent space. The potential drawback of these approaches is that the prior knowledge goes through an algorithm's internal processes, which may make it difficult to exactly understand how the prior knowledge affects the final output. As will be seen in Sect. 4, such an implicit, complex procedure prevents users from easily steering the clustering process and achieving the output that properly reflects their intent.

As for *flexibility*, many methods have been developed for handling specific situations built on particular assumptions about data. On the contrary, our method is capable of handling much broader situations involving various forms of prior knowledge generated by users, as discussed in Sect. 1. To be specific, WS-NMF/WS-SymNMF can solve traditional semi-supervised clustering problems where the partial pairwise constraints of data items are given. Rather than imposing a pairwise constraint form, which can possibly have a large number of different constraints with respect to the number of data items, our methods can also directly accommodate the hard-/soft-clustering membership coefficient as prior information. Finally, our methods can incorporate the information about the entire data.

3 Weakly supervised nonnegative matrix factorization

3.1 Preliminaries

Given a nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$ and an integer $k \ll \min(m, n)$, NMF finds a lower-rank approximation as

$$X \approx WH^T,$$

where $\mathbb{R}_+$ denotes the set of nonnegative real numbers, and $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{n\times k}$ are the two nonnegative factors. This lower-rank approximation problem can be formulated in terms of the Frobenius norm, i.e.,

$$\min_{W,\, H\geq 0} \left\| X - WH^T \right\|_F^2. \tag{1}$$

When NMF is used in clustering applications, the k columns of W are viewed as the representative vectors of k clusters (e.g., cluster centroids), and the n rows of H represent the (soft-) clustering membership coefficients of n individual data items. By setting the cluster index of each data item as the column index corresponding to the maximum value in each row vector of H, we obtain the hard-clustering results.

On the other hand, given an input matrix $S \in \mathbb{R}_+^{n\times n}$ in the form of a similarity matrix or an adjacency matrix of a graph, a symmetric NMF is formulated as

$$\min_{H\geq 0} \left\| S - HH^T \right\|_F^2, \tag{2}$$

where $H \in \mathbb{R}_+^{n\times k}$ is the factor matrix. In clustering, H is interpreted as the cluster membership coefficients similar to those in NMF,[1] and it performs similar to, or in some cases, better than well-known methods such as spectral clustering (Kuang et al. 2012, 2014).

3.2 Formulation

We now present the formulations of WS-NMF and symmetric NMF (WS-SymNMF) in clustering applications. In the following formulations, we assume that we are given particular prior knowledge about W and/or H shown in Eqs. (1) and (2). The prior knowledge is manifested in the form of reference matrices for W and H. These reference matrices play a role of making W and H become similar to them.

3.2.1 Weakly-supervised NMF (WS-NMF)

Given nonnegative reference matrices $W_r \in \mathbb{R}_+^{m\times k}$ for W and $H_r \in \mathbb{R}_+^{n\times k}$ for H and their nonnegative diagonal mask/weight matrices $M_W \in \mathbb{R}_+^{k\times k}$ and $M_H \in \mathbb{R}_+^{n\times n}$ along with an input data matrix $X \in \mathbb{R}_+^{m\times n}$ and an integer $k \ll \min(m, n)$, WS-NMF minimizes the following objective function with the additional regularization terms that penalize the differences between H_r and H (up to a row-wise scaling) and those between W_r and W,

[1] For the sake of notation simplicity, we do not distinguish H between Eqs. (1) and (2).

$$f(W,\ H,\ D_H) = \min_{W,\ H,\ D_H} \left\| X - WH^T \right\|_F^2 + \|(W - W_r)\, M_W\|_F^2 + \|M_H\,(H - D_H H_r)\|_F^2 \quad (3)$$

by finding nonnegative factors $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{n\times k}$ and a nonnegative diagonal matrix $D_H \in \mathbb{R}_+^{n\times n}$.

3.2.2 Weakly-supervised symmetric NMF (WS-SymNMF)

Given a nonnegative reference matrix $H_r \in \mathbb{R}_+^{n\times k}$ for H and its nonnegative diagonal mask/weight matrix $M_H \in \mathbb{R}_+^{n\times n}$ along with a symmetric nonnegative matrix $S \in \mathbb{R}_+^{n\times n}$ and an integer $k \ll \min(m,\ n)$, WS-SymNMF solves

$$\min_{H,\ D_H} \left\| S - HH^T \right\|_F^2 + \|M_H\,(H - D_H H_r)\|_F^2 \quad (4)$$

for a nonnegative factor $H \in \mathbb{R}_+^{n\times k}$ and a nonnegative diagonal matrix $D_H \in \mathbb{R}_+^{n\times n}$.

The original symmetric NMF algorithm (Kuang et al. 2012) has utilized a projected-Newton-based algorithm to solve the fourth-order optimization problem in Eq. (2). However, after adding the regularization term as shown in Eq. (4), the computation becomes much more expensive. To avoid this problem, we employ a recent improvement of the symmetric NMF formulation (Kuang et al. 2014), which decouples the product of a single factor H into that of the two different factors H_1 and H_2. In other words, we turn Eq. (4) into

$$g(H_1,\ H_2,\ D_H) = \min_{H_1,\ H_2,\ D_H} \left\| S - H_1 H_2^T \right\|_F^2 + \mu\, \|H_1 - H_2\|_F^2 + \frac{1}{2}\sum_{i=1}^{2} \|M_H\,(H_i - D_H H_r)\|_F^2\,, \quad (5)$$

where we enforce H_1 and H_2 to be similar to each other via the second term. This formulation significantly reduces the computational cost compared to solving Eq. (4) by a Newton-type method since the highly efficient block coordinate decent method developed for the standard NMF can be utilized (Kim and Park 2011; Kim et al. 2014). The detailed algorithm based on this formulation will be described in Sect. 3.3.2.

3.2.3 Interpretation

WS-NMF and WS-SymNMF enable users to impose various types of prior knowledge in the regularization terms in Eqs. (3) and (5).

First, the rows of H_r specify the prior information about the soft-clustering membership of individual data items using the third term in Eqs. (3) and (5). Note that our formulation contains a diagonal matrix D_H as a variable to optimize since D_H

can handle potentially different scales between H_r and H. For example, the two vectors, say, (0.1, 0.3, 0.6) and (0.2, 0.6, 1.2), are not different from each other as the cluster membership coefficients, and D_H allows us to ignore this scaling issue. On the other hand, users may not want to specify all the n rows (or n data items) in H_r, but instead, they may be interested in imposing their prior knowledge on a partial set of data items. The diagonal matrix M_H, placed on the left side of H_r and H, can deal with this situation by masking or down-weighting those rows or data items in H_r whose cluster membership coefficients are not to be regularized or supervised.

Second, the columns of W_r specify the cluster centroid/basis representations, as shown in the second term in Eq. (3). For example, in the case of document clustering, the columns of W_r correspond to the topic clusters typically represented by their most frequent keywords, and users may want to manipulate these keywords to properly steer the semantic meaning of the topic. When users want to specify only a subset of clusters in W_r instead of the entire k clusters, the diagonal matrix M_W, placed on the right side of W_r, plays the role of masking or down-weighting those columns or cluster centroids in W_r that are to be ignored or considered as less important than the others. Unlike H_r, the regularization term for W_r does not involve any additional diagonal matrix analogous to D_H, which could adjust the scale of W to that of W_r. This is because it is sufficient to handle the scaling of only one of the two NMF factors, W and H, due to the relationship

$$WH^T = WDD^{-1}H^T = (WD)\left(HD^{-1}\right)^T,$$

which implies that if W and H are the solution of a particular NMF problem, then so are WD and HD^{-1} for any element-wise positive diagonal matrix D.

Finally, note that our formulations do not have typical regularization parameters that assign different weights on each term because such weighting is carried out by M_W and M_H. For instance, assuming equal weights on each column of W_r and on each row of H_r, M_W and M_H can be simplified as

$$M_W = \alpha I_k \quad \text{and} \quad M_H = \beta I_n, \tag{6}$$

respectively, where $I_k \in \mathbb{R}^{k\times k}$ and $I_n \in \mathbb{R}^{n\times n}$ are identity matrices. Applying Eq. (6) to Eqs. (3) and (5), we obtain

$$\begin{aligned}
&\min_{W,\,H,\,D_H} \left\| X - WH^T \right\|_F^2 + \alpha \left\| W - W_r \right\|_F^2 + \beta \left\| H - D_H H_r \right\|_F^2 \text{ and} \\
&\min_{H_1,\,H_2,\,D_H} \left\| S - H_1 H_2^T \right\|_F^2 + \mu \left\| H_1 - H_2 \right\|_F^2 + \frac{1}{2}\beta \sum_{i=1}^{2} \left\| H_i - D_H H_r \right\|_F^2,
\end{aligned} \tag{7}$$

which are controlled by scalar regularization parameters α and β.

3.3 Algorithm

Our algorithms to solve WS-NMF and WS-SymNMF are based on the block coordinate descent framework. Basically, we divide the entire variables into several subsets, e.g., W, H, and D_H in Eq. (3), and H_1, H_2, and D_H in Eq. (5), respectively. Then, we iteratively solve for each subset of variables at a time while fixing the remaining variables. Each sub-problem can then be formulated as a nonnegativity-constrained least squares (NLS) problem except for D_H, which has a closed form solution at each iteration. In the following, we describe our algorithm details for WS-NMF and WS-SymNMF.

3.3.1 Weakly-supervised NMF (WS-NMF)

To solve Eq. (3), we iteratively update W, H, and D_H as follows. Assuming the initial values of these variables are given, we update W by solving an NLS problem as

$$\min_{W \geq 0} \left\| \begin{bmatrix} H \\ M_W \end{bmatrix} W^T - \begin{bmatrix} X^T \\ M_W W_r^T \end{bmatrix} \right\|_F^2. \tag{8}$$

Next, we update each row of H, i.e., $H(i, :)$, one at a time by solving another NLS problem as

$$\min_{H \geq 0} \left\| \begin{bmatrix} W \\ M_H(i)\, I_k \end{bmatrix} H(i, :)^T - \begin{bmatrix} X(:, i) \\ M_H(i)\, D_H(i) H_r(i, :)^T \end{bmatrix} \right\|_F^2, \tag{9}$$

Finally, we update the ith diagonal component $D_H(i)$ of D_H as

$$D_H(i) = \begin{cases} \frac{H_r(i,:) \cdot H(i,:)^T}{\|H_r(i,:)\|_2^2} & \text{if } M_H(i) \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

3.3.2 Weakly-supervised symmetric NMF (WS-SymNMF)

To solve Eq. (5), we iteratively update H_1, H_2, and D_H as follows. Assuming the initial values of these variables are given, we update H_1 by solving an NLS problem as

$$\min_{H_1 \geq 0} \left\| \begin{bmatrix} H_2 \\ \sqrt{\mu} I_k \\ \frac{M_H(i)}{\sqrt{2}} I_k \end{bmatrix} H_1(i, :)^T - \begin{bmatrix} S(:, i) \\ \sqrt{\mu} H_2(i, :)^T \\ \frac{M_H(i) D_H(i)}{\sqrt{2}} H_r(i, :)^T \end{bmatrix} \right\|_F^2. \tag{11}$$

Next, we update H_2 in a similar manner by solving

$$\min_{H_2 \geq 0} \left\| \begin{bmatrix} H_1 \\ \sqrt{\mu} I_k \\ \frac{M_H(i)}{\sqrt{2}} I_k \end{bmatrix} H_2(i, :)^T - \begin{bmatrix} S(:, i) \\ \sqrt{\mu} H_1(i, :)^T \\ \frac{M_H(i) D_H(i)}{\sqrt{2}} H_r(i, :)^T \end{bmatrix} \right\|_F^2. \tag{12}$$

Finally, we update the ith diagonal component $D_H(i)$ of D_H as

$$D_H(i) = \begin{cases} \frac{H_r(i,:)\cdot(H_1(i,:)+H_2(i,:))^T}{2\|H_r(i,:)\|_2^2} & \text{if } M_H(i) \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

3.3.3 Discussions

Nonnegativity-constrained least squares. NLS problems such as Eqs. (8), (9), (11), and (12) play crucial roles in the proposed algorithms. We adopt a recently proposed block principal pivoting (BPP) algorithm for NLS problems (Kim and Park 2011), which is known to be one of the fastest algorithms. The BPP algorithm is basically an active-set type of constrained optimization method, which solves these least squares problems under nonnegativity constraints. By aggressively changing a number of variables in the active set at each iteration instead of changing each of them one at a time, it achieves significant improvement in computational efficiency over classical active-set algorithms for NLS.

One potential drawback that may slow down the algorithms of WS-NMF/WS-SymNMF when using the BPP algorithm is that we cannot exploit the structural aspect of having a common left-hand side matrix for multiple right-hand side matrices in the least-squares formulation, as originally presented in Kim and Park (2011). That is, in Eqs. (9), (11), and (12), the left-hand side matrices multiplied by $H(i,:)^T$, $H_1(i,:)^T$, and $H_2(i,:)^T$, respectively, do not remain the same with respect to the value of i while the matrix multiplied by W^T in Eq. (8) does throughout all the rows of W. The original BPP algorithm exploits this common left-hand side matrix by re-using the matrix multiplications commonly used for solving the multiple right-hand side cases. Nonetheless, the only difference in the lefft-hand side matrices in our formulation is the different coefficients corresponding to an identity matrix I_k, e.g., $M_H(i)$ in Eq. (9) and $\frac{M_H(i)}{\sqrt{2}}$ in Eqs. (11) and (12). Therefore, by performing the simple additional step of k additions to the diagonal entries of the common left-hand side matrices, we can still take full advantage of the efficient BPP algorithm.

Convergence and stopping criteria. The block coordinate descent framework that we adopt guarantees that our algorithms converge to a stationary point as long as the unique global solution can be obtained for each sub-problem (Bertsekas 1999). Obviously, all the sub-problems shown in Eqs. (8), (9), (11), and (12) are strongly convex since all of them have second-order objective functions with linear constraints, and thus the global solution for each sub-problem can be computed, which ensures the convergence of our algorithm.

The stopping criterion for our algorithm is to check whether the projected gradient values (Lin 2007) of Eqs. (3) and (5) become zero, indicating that the algorithm reached a stationary point. Specifically, considering potential numerical errors, we use the stopping criteria for WS-NMF and WS-SymNMF as

$$\frac{\left\|\nabla^P f(W, H, D_H)\right\|_F}{\left\|\nabla^P f(W^{(1)}, H^{(1)}, D_H^{(1)})\right\|_F} \leq \epsilon \quad \text{and} \quad \frac{\left\|\nabla^P g(H_1, H_2, D_H)\right\|_F}{\left\|\nabla^P g(H_1^{(1)}, H_2^{(1)}, D_H^{(1)})\right\|_F} \leq \epsilon,$$

respectively, where the denominators represent the Frobenius norm of the gradient evaluated at the output obtained at the first iteration.

Initialization. Since the WS-NMF/WS-SymNMF formulations are non-convex, we obtain only a local optimum, which may change depending on the initialization. In WS-NMF/WS-SymNMF, for those columns of W and rows of H whose corresponding reference information are given via W_r and H_r (along with corresponding non-zero diagonal entries in M_W and M_H, respectively), we set their initial values to be the same as those in W_r and H_r. For the rest, we set them as the values uniformly sampled between zero and the maximum element in the already initialized sub-matrices in the previous step. When M_W and/or M_H are zero matrices, we set each element of W and H as the value uniformly sampled between zero and the maximum element of the input matrix X or S.

Computational complexity. The complexities of WS-NMF and WS-SymNMF are comparable to those of NMF and SymNMF since all of them follow the similar block coordinate descent framework. In the WS-NMF and WS-SymNMF algorithms, except for D_H in Eqs. (10) and (13), which are computed relatively fast, all the sub-problems have more rows in both the left- and right-hand sides in Eqs. (8), (9), (11), and (12) compared to the standard NMF algorithm. However, we convert the original least squares problems (e.g., $\min_x \|Ax - b\|_2$) into their normal equations (e.g., $A^T Ax = A^T b$) before solving these sub-problems. In both our proposed methods and the standard NMF, those matrices corresponding to $A^T A$ have the same size of $k \times k$. Thus, the complexity of NLS problems, which comprise a majority of computational time taken in WS-NMF/WS-SymNMF, remains unchanged compared to the standard NMF.

Parameter selection. WS-NMF/WS-SymNMF methods contain several parameters to tune. First, as for the parameter μ in Eq. (5), we found that the final output is insensitive to its value. In fact, the only difference between the sub-problems of H_1 and H_2, Eqs. (11) and (12), is H_2 in the place of H_1 (and vice versa) in the left-hand and the right-hand sides. In this respect, as they get closer to each other, their sub-problems naturally become more similar, resulting in almost identical solutions for H_1 and H_2. Based on this observation, we set μ as the squared value of the maximum entry of S, i.e.,

$$\mu = \max_{i,j} \left(S_{ij}\right)^2,$$

so that the scale of the second term is comparable to that of the first term in Eq. (5).

Second, as for the weighting matrices M_W and M_H in Eqs. (3) and (5), we can fine-tune their non-zero diagonal values by analyzing how strongly to impose the prior information in the final solution. In practice, if the cluster membership based on H deviates too much from that based on H_r, we can increase the diagonal entries of M_H. When regularizing the basis vectors using W_r, we can use the same strategy. Alternatively, we can measure the difference between W and W_r in terms of their Frobenius norms or the ranking of their feature importance values, e.g., the most representative terms in document clustering, as will be discussed in the next section.

Table 1 Summary of data sets used in the semi-supervised clustering experiments

	Document data				Facial image data	
	20News	RCV1	NIPS	Cora	ExtYaleB	AR
No. dimensions	13,656	10,284	17,583	20,110	3,584	4,800
No. data	943	1,210	420	573	2,414	2,600
No. clusters	20	40	9	20	38	68

4 Experiments

To evaluate the performance of our approaches, we conduct both quantitative and qualitative analyses. For the quantitative analysis, we perform standard semi-supervised clustering experiments, and for the qualitative analysis, we present several user-driven clustering scenarios using WS-NMF/WS-SymNMF.

4.1 Semi-supervised clustering

In this experiment, we conduct standard semi-supervised clustering experiments in which we evaluate the improvement in clustering performance given partial label information. We used four widely-used document data sets: 20 Newsgroups (20News),[2] RCV1-v2 (RCV1),[3] NIPS1-17 (NIPS),[4] and Cora,[5] and two facial image data sets: extended Yale face database B (ExtYaleB)[6] and AR face database (AR).[7] These data sets are summarized in Table 1. All of them are encoded as high-dimensional vectors using a bag-of-words encoding scheme for document data and rasterized pixel intensity values for facial image data. Their class labels indicate their associated topic categories for document data and person ID's for facial image data. All the data sets are normalized so that each data vector has a unit L_2-norm, which is a common practice when NMF is applied in clustering (Kim and Park 2008; Kuang et al. 2012).

For comparisons, we selected two semi-supervised clustering methods based on NMF: the one proposed by Chen et al. (2008) (SS-NMF1) and the other proposed by Lee et al. (2010) (SS-NMF2). In addition, we also included three well-known semi-supervised clustering methods: metric pairwise constrained k-means (MPCK-Means) (Bilenko et al. 2004), pairwise constrained k-means (PCK-Means) (Bilenko et al. 2004), and semi-supervised spectral clustering (SS-Spectral) (Kulis et al. 2009) as well as two other baseline clustering methods without any supervision: k-means

[2] http://qwone.com/~jason/20Newsgroups/.

[3] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm .

[4] http://ai.stanford.edu/~gal/data.html.

[5] http://people.cs.umass.edu/~mccallum/data.html.

[6] http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html.

[7] http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html.

and NMF. For SS-NMF1, MPCK-Means, and PCK-Means, we used their publicly available source codes.[8,9]

We randomly selected a particular number of data items along with their labels for the semi-supervised portion. For WS-NMF in Eq. (3), we set $M_W = 0$ (no supervision on the basis vector matrix W) and set M_H and H_r as

$$M_H(i) = \begin{cases} \beta & \text{if the } i\text{th data item is supervised} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

and

$$H_r(i,\ j) = \begin{cases} 1 & \text{if the label of the } i\text{th data item is } j \\ 0 & \text{otherwise} \end{cases}.$$

In Eq. (14), β controls how strongly the supervision is imposed. In order to avoid too strong or weak supervision, we start from a sufficiently large value and then exponentially decrease β until the clustering accuracy for the supervised data becomes worse than 95 %. Although this approach does not satisfy all the given constraints, it generally led us to reasonable clustering performance on unseen data in practice.

For SS-NMF1, we need to construct a similarity matrix. To this end, we used a Gaussian kernel $K(x_1,\ x_2) = \exp\left(-\|x_2 - x_1\|^2 / 2\sigma^2\right)$ with a bandwidth parameter $\sigma = 10^{-5}$, as suggested in Chen et al. (2008). For SS-NMF2, we implemented the method with the same parameter setting as described in Lee et al. (2010).

MPCK-Means and PCK-Means require pairwise constraints, e.g., must-links or cannot-links. To convert the labels of supervised data to their pairwise constraints, we generated the pairwise constraints as follows.

1. For each cluster, assuming that n_i supervised data items are given for cluster i, we generate $(n_i - 1)$ must-link constraints between the (1, 2), (2, 3), ..., $(n_i - 1,\ n_i)$th data items, respectively.
2. Given k clusters in total, we generate a single cannot-link constraint between a randomly selected pair of data items from clusters i and $(i + 1)$, where $i = 1, 2, \ldots, (k - 1)$.

The set of pairwise constraints generated in this manner fully contains the label information of the given semi-supervised portion of data.

To run SS-Spectral, we first constructed the k-nearest neighbor graph among data items, G^X, where the edge weights correspond to their similarity values. Specifically, the edge weight in G^X between x_i and x_j is defined (Zelnik-Manor and Perona 2004) as

$$G^X_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{\sigma_i \sigma_j}\right) I\left(x_i \in N_l\left(x_j\right) \quad \text{or} \quad x_j \in N_l\left(x_i\right)\right),$$

[8] SS-NMF1: http://www-personal.umich.edu/~chenyanh/SEMI_NMF_CODE.zip.

[9] MPCK-Means and PCK-Means: http://www.cs.utexas.edu/users/ml/risc/code/.

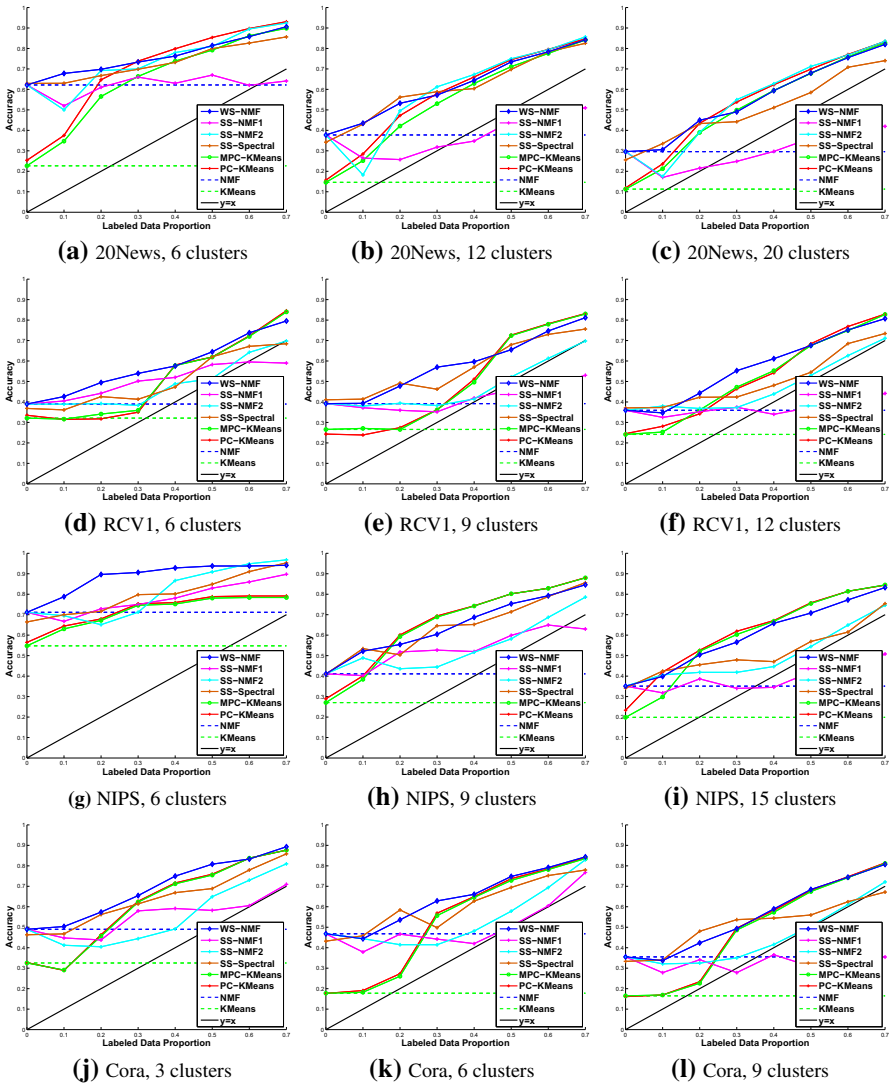
where σ_i is the distance between x_i and its lth nearest neighbor data item, e.g., $l = 7$ in our case (Zelnik-Manor and Perona 2004), and $I(\cdot) = 1$ if the condition in the parenthesis is true and zero otherwise, and $N_l(x_i)$ (or $N_l(x_j)$) is the set of the l nearest neighbors of x_i (or x_j). Next, every element in G^X is divided by the maximum value among all the entries in G^X so that all the elements of G^X lie between zero and one. Afterwards, we set those entries of G^X corresponding to must-links to one while setting those corresponding to cannot-links to zero. A re-normalization on the graph is then performed, followed by a typical spectral clustering process (Zelnik-Manor and Perona 2004).

Results and discussions. Figures 1 and 2 show the clustering accuracy depending on various supervised proportions of data. Between the two baseline methods, NMF shows superior performances compared to k-means. Starting from the baseline performances, all the six semi-supervised methods, WS-NMF, SS-NMF1, SS-NMF2, MPCK-Means, PCK-Means, and SS-Spectral, generally perform better as the level of supervision increases.

Compared to WS-NMF, SS-NMF1 and SS-NMF2 show relatively poor performances in most cases. Given a low level of supervision, e.g., 10–20 %, SS-NMF1 and SS-NMF2 often perform worse than unsupervised cases, as shown in Figs. 1a–c, g, j and 2d–f. Even with a high level of supervision, their performances do not significantly improve, either. For instance, the SS-NMF1 results are shown below the $y = x$ line in many cases, indicating that the given constraints are not even fully satisfied. As briefly discussed in Sect. 2, the reason is because of the complicated procedure of joint factorization and an additional k-means step on the NMF output, rather than directly interpreting the NMF output as the clustering labels. On the other hand, SS-NMF2 satisfies most of the given constraints, showing the performance lines well over the $y = x$ line, but as the level of supervision increases, the performance difference with respect to the $y = x$ line becomes small, as shown in Figs. 1d–f, i, l and 2a–f. It indicates that the main procedure of SS-NMF2, which imposes the clustering constraints to the input similarity matrix, can significantly distort the original data relationships, and thus it may not be able to reveal the inherent clusters in the entire data.

The two semi-supervised k-means methods, MPCK-Means and PCK-Means, show similar performance patterns to WS-NMF, but their performances sometimes stay at relatively low accuracy values even with a nontrivial level of supervision, e.g., 0–20 % of the data for supervision in the case of document data shown in Fig. 1d–f, k, l. In the case of image data shown in Fig. 2, the performance of PCK-Means becomes much worse than MPCK-Means, which indicates that without a metric learning step, the underlying nonlinearity in the image data cannot be properly handled by PCK-Means.

The semi-supervised spectral clustering method, SS-Spectral, shows relatively good performances compared to the other semi-supervised methods. Specifically, SS-Spectral works reasonably well with a low level of supervision, as shown in Figs. 1b, c, e, k, l and 2d–f, but its performance does not increase as much as the increased level of supervision, sometimes close to or even much worse than the $y = x$ line, as shown in Figs. 1d, f, i, l and 2a, b. It indicates that SS-Spectral may not be able to properly satisfy a large amount of given constraints. This could be because the constraints



(a) 20News, 6 clusters **(b)** 20News, 12 clusters **(c)** 20News, 20 clusters

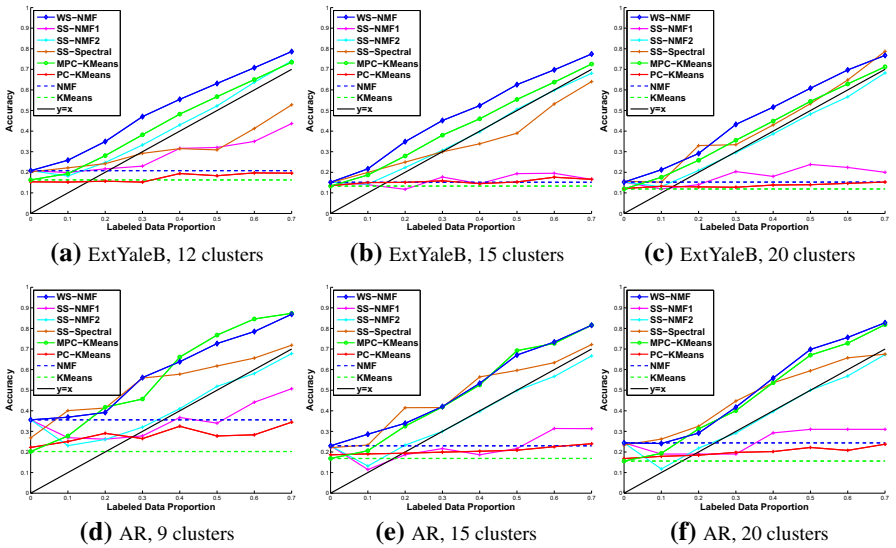
(d) RCV1, 6 clusters **(e)** RCV1, 9 clusters **(f)** RCV1, 12 clusters

(g) NIPS, 6 clusters **(h)** NIPS, 9 clusters **(i)** NIPS, 15 clusters

(j) Cora, 3 clusters **(k)** Cora, 6 clusters **(l)** Cora, 9 clusters

Fig. 1 Comparison of clustering accuracy for four document data sets summarized in Table 1. All the results are the averaged values of five randomly selected sets of clusters. A reference line, $y = x$, is drawn to show the amount of accuracy improvement for each method using partial supervision. In addition, the results of the two unsupervised methods, NMF and k-means, are also presented using *dashed horizontal lines*

of SS-Spectral are imposed into the input graph matrix by maximizing/minimizing the edge weights of must-links/cannot-links. However, such a supervision scheme at the level of an input matrix instead of direct constraints on the ouptut cluster labels does not necessarily guarantee that the clustering results generated by the subsequent process of SS-Spectral satisfy the given constraints.



(a) ExtYaleB, 12 clusters **(b)** ExtYaleB, 15 clusters **(c)** ExtYaleB, 20 clusters

(d) AR, 9 clusters **(e)** AR, 15 clusters **(f)** AR, 20 clusters

Fig. 2 Comparison of clustering accuracy for two facial image data sets summarized in Table 1. All the results are averaged values of five randomly selected sets of clusters. A reference line, $y = x$, is drawn to show the amount of accuracy improvement for each method using partial supervision. In addition, the results of the two unsupervised methods, NMF and k-means, are also presented using *dashed horizontal lines*

On the other hand, WS-NMF does not show the above-mentioned undesirable behaviors of the other semi-supervised methods. At the same time, its performance is shown to be consistently better than the other semi-supervised methods over a wide range of the supervision level. Another point to note is that the performance gaps between WS-NMF and other methods are the most significant especially when the level of supervision is relatively small up to around 20–30 %. Such an observation is particularly important because in reality, only a small level of supervision is affordable, say, much less than 20–30 %, which makes WS-NMF a promising technique in practical semi-supervised clustering applications.

4.2 User-driven clustering scenarios

In this section, we show three user-driven clustering scenarios using our methods. The first scenario shows the case of weakly imposing the reference information easily obtained about the data in WS-NMF so that the clustering results can reflect such information. The second scenario presents the case where WS-SymNMF improves the initial clustering results via partial supervision on user-selected data. The last scenario demonstrates the case where the reference information about the basis vectors is imposed in WS-NMF to steer the clustering results in a semantically meaningful manner.

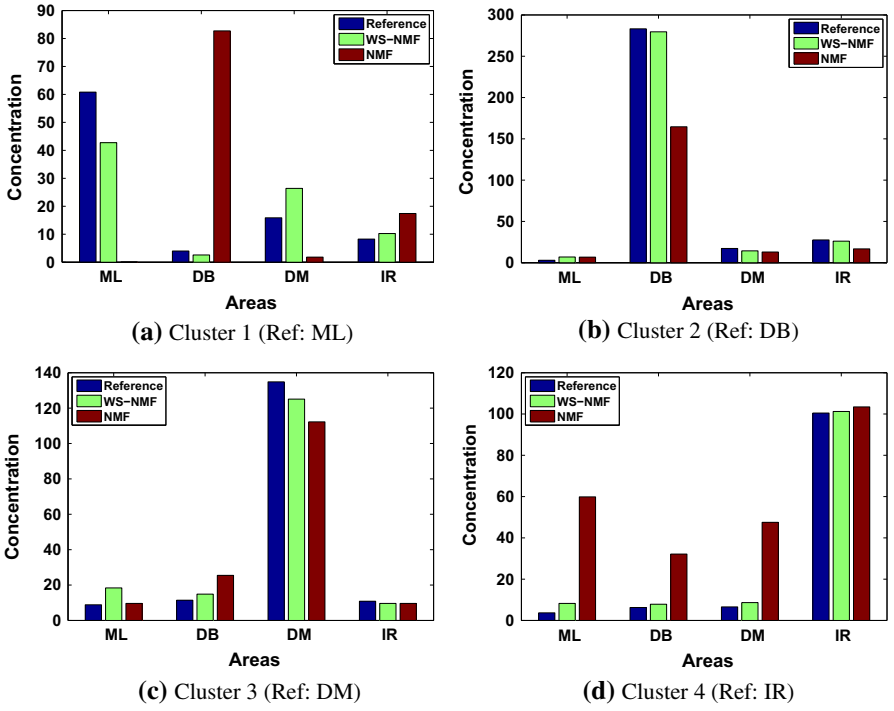
To show our clustering scenarios, we selected two real-world data sets: *Four Area* and *IMDB* (Gupta et al. 2012).[10] *Four Area* data set is the collection of papers published in machine learning (ML), databases (DB), data mining (DM), and information retrieval (IR) ranging from years 2001 to 2008. The data set contains various types of information about a paper such as a title, an author list, a venue, and a publication year. Each of the four areas contains the papers published in five major conference venues in each field. From this data set, we represent each author as a collection of papers he/she wrote, which is then encoded as a bag-of-words expression using the paper title field. In this manner, we obtain an author-by-term matrix whose element indicates the frequency of the corresponding term appearing in the papers that the corresponding author wrote. By selecting 703 authors who wrote the most number of papers, the final matrix has the size of $703 \times 2{,}361$, encoded using 2,361 distinct terms.

IMDB data set is the movie collection released from years 2006 to 2009. It contains the following information about movies: a title, an actor/actress list, and multi-labeled genres. From four genres, animation, horror, music show, and action, we construct a co-starring graph between actors/actresses in which the edge values are specified as the number of movies that co-starred the two actors/actresses. From this co-starring graph, we sample 166 actors/actresses with the most movies, and finally, the resulting graph based on 2,092 movies is normalized in the form of $D^{-1/2}SD^{-1/2}$ where S is the co-starring graph and D is its degree matrix.

Scenario 1: weak supervision via readily available information. In this experiment, we show the usage scenario of WS-NMF to perform clustering on *Four Area* data set. As reference information, we utilize venue information, which is readily available from the data set. The intent of imposing such prior knowledge in the clustering process is to avoid drastically different clustering results from the groupings based on the venue information, but at the same time, we still want to identify some outlying authors who publish in a certain area (or a venue) those papers having unusual topics. To this end, we form a four-dimensional vector for each author by counting the number of papers he/she wrote in each of the four areas solely based on the venue information. We set this vector as the corresponding row vector in H_r. For example, if an author wrote three papers in the conferences of the ML category and two papers in the DM category, the corresponding row vector in H_r would be $(2,\ 0,\ 0,\ 3)$, where the four dimensions are in an order of DM, DB, ML, and IR categories. Note that the reference matrix H_r does not need to be normalized because WS-NMF automatically adjusts it via D_H in Eq. (3).

We assign equal weights in M_H to all data items, i.e., $M_H = \beta I_n$. We determine β based on how many (hard-)cluster memberships are different between H_r and H. The (hard-)cluster membership is obtained as the row index whose entry is the largest in the corresponding row of H_r and H. In general, as β increases, H is enforced more strongly to be close to H_r, resulting in fewer differences in the cluster membership. In this experiment, we set β such that the number of cluster membership differences between H_r and H is about 12 % of the total data items. On the other hand, when running the standard NMF, about 43 % of data items are shown to change their cluster memberships from the reference matrix H_r.

[10] http://dais.cs.uiuc.edu/manish/ECOutlier/.



(a) Cluster 1 (Ref: ML) **(b)** Cluster 2 (Ref: DB)

(c) Cluster 3 (Ref: DM) **(d)** Cluster 4 (Ref: IR)

Fig. 3 Concentrations of four areas in each cluster generated by (1) reference information only, (2) WS-NMF with reference information, and (3) the standard NMF for *Four Area* data set

Now we have three different clustering results obtained by (1) the reference data (solely based on H_r), (2) WS-NMF, and (3) the standard NMF. Figure 3 shows how these three cases differ in terms of the concentrations of the four areas in each cluster. To obtain the results shown in this figure, we first normalize each row of H_r so that their row sum equals to one. Next, for all the data items in each cluster from the three different clustering results, we sum up their rows. As a sanity check, one can see that each cluster obtained by the reference data (blue color) always shows the maximum value in its corresponding area since each vector contributing to a particular cluster always has the maximum entry in the corresponding cluster. In contrast, in the clustering results of the standard NMF (brown color), which does not take into account the reference information, the concentration of the DB area in cluster 2 is relatively small compared to the two other approaches while cluster 1 now shows the highest concentration in the DB area among the four areas, which results in both clusters 1 and 2 representing the DB area.

Table 2 shows the most frequently used keywords in each cluster for the three different clustering results. While Fig. 3 is about the cluster quality in terms of how much each of the four areas is separated/mixed in the resulting clusters, Table 2 tells us how clear or distinct the topical meaning of each cluster is. From these keyword summaries of clusters, clusters 1 and 2 in the standard NMF turn out to overlap with each other, as shown in the keywords such as ‘queri’, ‘index’, ‘effici’, and ‘databas’,

whereas these two clusters in the reference data show clearly distinct topics with no overlapping keywords except for generic terms such as 'data'. This observation indicates that clusters 1 and 2 generated by the standard NMF are not clearly separated in terms of either their topics or area concentrations.

On the contrary, WS-NMF shows similar behaviors to the reference data in both Table 2 and Fig. 3. Our analysis on individual authors reveals that their cluster membership differences shown in the WS-NMF result mostly make sense. For instance,

Table 2 Representative keywords of each cluster for *Four Area* data set

Ref.	WS-NMF	NMF
Cluster 1 (Ref: ML)		
learn	learn	xml
model	cluster	queri
data	data	data
algorithm	classif	index
supervis	analysi	effici
reinforc	text	system
kernel	reinforc	xqueri
base	algorithm	relat
space	model	document
inform	multi	databas
Cluster 2 (Ref: DB)		
queri	queri	data
xml	xml	queri
data	data	stream
databas	databas	databas
stream	stream	system
effici	effici	effici
index	index	process
system	system	index
join	join	manag
web	process	join
Cluster 3 (Ref: DM)		
mine	mine	mine
data	data	data
cluster	cluster	pattern
pattern	pattern	cluster
frequent	frequent	frequent
associ	associ	associ
base	base	base
rule	rule	effici
effici	effici	rule
algorithm	algorithm	set

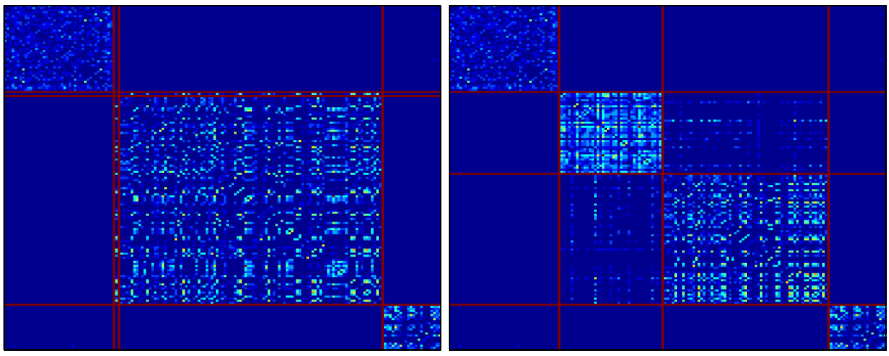
Table 2 continued

Ref.	WS-NMF	NMF
Cluster 4 (Ref: IR)		
retriev	retriev	web
web	web	learn
queri	queri	retriev
document	document	text
search	inform	cluster
inform	search	inform
base	base	base
text	model	data
model	languag	model
languag	text	document

one can see that cluster 4 (Ref: IR) has larger concentrations of the ML, DB, and DM categories in WS-NMF compared to the reference data. We found that some of the authors with their cluster membership changes to cluster 4 in the WS-NMF result are D. Sivakumar and A. G. Hauptmann. That is, D. Sivakumar, who is labeled as cluster 2 (Ref: DB) by the reference data but as cluster 4 by WS-NMF, published the papers of the IR topic in the DB venues, e.g., 'comparing and aggregating rankings with ties' in PODS and 'self-similarity in the web' in VLDB. A. G. Hauptmann, who is labeled as cluster 1 (Ref: ML) by the reference data but as cluster 4 by WS-NMF, published the papers of the IR topic in the ML venues, e.g., 'learning to select good title words: a new approach based on reverse information retrieval' in ICML.

In addition, between cluster 1 (Ref: ML) and cluster 3 (Ref: DM), the cluster membership changes due to WS-NMF include reasonable examples. For instance, C. Ding, who is labeled as cluster 3 by the reference data but as cluster 1 by WS-NMF, wrote the papers in the DM venues mostly about dimension reduction and clustering techniques that heavily involve theoretical formulations and algorithms. C. Elkan, who is labeled as cluster 1 by the reference data but as cluster 3 by WS-NMF, wrote the papers closely related to applications in the ML venues, e.g., 'Bayesian approaches to failure prediction for disk drives' in ICML.

Scenario 2: semi-supervision with exemplars. For this experiment, we use *IMDB* data set and show a semi-supervised clustering scenario via partial label information. Since the data set is a co-starring graph, we use WS-SymNMF with rank $k = 4$ and compare its performance with SymNMF. Figure 4a shows the clustering results of SymNMF, where the clusters are significantly unbalanced, e.g., 42, 1, 101, and 22. We now look into cluster 3 because it is the largest cluster, which is usually difficult to understand due to its size and noisiness. We then find that the members in cluster 3 are mostly involved in music shows all over the world. To further divide it into meaningful groups, we sample five actors/actresses appearing mostly in the music shows held in the U.S. and another five in the other countries such as those in Europe. These samples include G. Strait in the U.S., an American country singer, and A. Montero, a famous singer in Spain. Now, we assign their corresponding row vectors of the reference matrix H_r in Eq. (5) as $(0,\ 1,\ 0,\ 0)$ for those actors/actresses in the U.S. and $(0,\ 0,\ 1,\ 0)$ for those



(a) SymNMF **(b)** WS-SymNMF

Fig. 4 Effects of semi-supervision in WS-SymNMF for *IMDB* data set

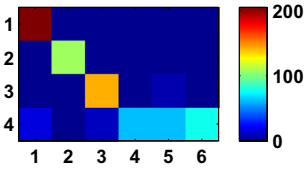
Table 3 Cluster quality measures for *IMDB* data set

	Total cut	Normalized cut	Lower-rank approx. error
WS-SymNMF	644	0.339	108.959
SymNMF	204	1.039	83.113
Spectral clustering	313	0.5778	N/A

in the other countries. We set their corresponding diagonal entries in M_H as one and the rest of the actors/actresses as zero in order to impose the label information only on the sample data.

The clustering results obtained by running WS-SymNMF with the above-described reference information are shown in Fig. 4b. Aside from clusters 1 and 4, which remain unchanged, one can see that WS-SymNMF divides cluster 2 in the SymNMF result into two well-separated clusters. As we check individual actors/actresses in these two clusters, those in the first cluster are shown to be mostly U.S. singers while those in the second cluster are European singers. One can also see that Fig. 4b shows some edges with large weight values between the two clusters, i.e., the bright-colored elements in the (2, 3)th and (3, 2)th sub-matrices. These actors/actresses turn out to be the top famous singers actively appearing in both the U.S. and European music shows, e.g., C. Aguilera, M. J. Blige, and 50 Cents in 'Grammy Awards', 'MTV Europe Music Awards', and 'Brit(ish) Awards'. We conjecture that these highly active singers across the world are the main cause of these two clusters being merged into one in the original SymNMF result.

Finally, we compare the cluster quality by using several widely-used measures, as shown in Table 3. Given a graph S, the total and the normalized cuts are defined as $\sum_{i=1}^{n}\sum_{j=1}^{n} S(i,\ j)\, I\left(c_i \neq c_j\right)$ and $\sum_{i=1}^{n} \frac{\sum_{j=n}^{n} S(i,\, j) I\left(c_i \neq c_j\right)}{\sum_{j=1}^{n}\sum_{l=1}^{n} S(j,\, l) I\left(c_j = c_i\right)}$, respectively, where c_i (or c_j) is the cluster index of the ith (or jth) data item. These measures represent the total sum of the inter-edge values between clusters, and thus a lower value indicates a better clustering quality. Finally, the lower-rank approximation error we used refers to the original SymNMF objective function value, i.e., $\|S - HH^T\|_F$.



(a) Confusion matrix

Clusters		
4	5	6
classif	cluster	retriev
learn	web	web
text	learn	search
web	data	inform
structur	analysi	queri
base	text	document

(b) Representative keywords of clusters 4, 5, and 6

Fig. 5 Effects of semi-supervision on basis vectors in WS-NMF for *Four Area* data set. The (i, j)th component of the confusion matrix indicates the number of authors that moved from cluster i to cluster j due to semi-supervision

In addition to the results of SymNMF and WS-SymNMF, Table 3 includes the standard spectral clustering results (Shi and Malik 2000) as baseline measures.

As shown in this table, WS-SymNMF degrades the original objective function value of SymNMF because of an additional regularization term included in WS-SymNMF. WS-SymNMF is also shown to be worse than SymNMF in terms of the total cut because the total cut generally favors unbalanced clusters. However, in terms of the normalized cut, which does not have this artifact and thus is widely used in graph clustering, WS-SymNMF shows a better result than SymNMF. Furthermore, WS-SymNMF performs even better than spectral clustering, which directly optimizes the normalized cut measure. This indicates that by incorporating a small portion of prior information (about 6 % in our case), WS-SymNMF can lead to both quantitatively and semantically better clustering results.

Scenario 3: semi-supervision on basis vectors. In this scenario, we showcase semi-supervision on basis vectors, which is one of the unique capabilities of WS-NMF compared to other semi-supervised approaches. We focus on cluster 4 generated by NMF for *Four Area* data set, and as shown in Table 2 and Fig. 3, cluster 4 is closely related to the IR area/topic that mainly handles text documents or web data. Inspired by the most frequent keywords generated by NMF, we decide to further explore this cluster from the three perspectives: (1) classification, (2) clustering, and (3) search/retrieval, which can be considered as general sub-topics in IR. In order to achieve this task, we manipulate the matrix $W \in \mathbb{R}_+^{2361\times 4}$ obtained by the standard NMF in Eq. (1), as follows. We first replicate the fourth column corresponding to the IR cluster twice, concatenate them on the right side of W, and set it as an initial reference matrix W_r for WS-NMF in Eq. (3). Next, for the first three columns corresponding to non-IR clusters in W_r, we set the values corresponding to IR-specific terms such as 'document', 'text', and 'web' as zero, which will discourage the documents containing these words to be clustered to the first three clusters. For the last three columns that have all the same vectors, we double the value corresponding to the term 'classification' in the first column, 'clustering' in the second, and 'retrieval' in the third so that we can steer these three topics as distinct clusters with our intended meanings.

Figure 5 summarizes the results obtained by running WS-NMF with an adjusted rank $k = 6$ and the above-described reference matrix W_r. As shown in Fig. 5a, very

few cluster membership changes are made among the first three non-IR clusters. On the other hand, those authors in cluster 4 in NMF move almost evenly to clusters 4, 5, and 6 after performing WS-NMF. Figure 5b confirms that these three clusters are indeed formed in our intended manner showing the terms, 'classif', 'cluster', and 'retriev', as the most frequently used terms in each cluster, respectively. The authors in these clusters include B. Zhang (cluster 4), who wrote the papers 'improving text classification using local latent semantic indexing' and 'web page classification through summarization', I. S. Dhillon (cluster 5), who wrote the papers 'iterative clustering of high-dimensional text data augmented by local search' and 'co-clustering documents and words using bipartite spectral graph partitioning', and W. B. Croft (cluster 6), who wrote the papers 'evaluating high accuracy retrieval techniques' and 'answer models for question answering passage retrieval'.

Overall, the three scenarios we presented in this section clearly show the extensive applicability of WS-NMF/WS-SymNMF by incorporating various types of prior knowledge in the clustering processes. In this manner, our methods allow us to obtain semantically meaningful clusters as well as crucial insights about data.

5 Conclusions and future work

In this paper, we presented novel methods called weakly supervised NMF (WS-NMF) and symmetric NMF (WS-SymNMF), which flexibly support user-driven clustering processes by incorporating various types of prior knowledge. Our contributions include (1) the formulation and the algorithm of WS-NMF/WS-SymNMF, (2) quantitative comparisons against well-known existing methods, and (3) user-driven clustering scenarios using different kinds of prior knowledge.

As seen from several real-world usage scenarios, WS-NMF/WS-SymNMF has a great potential to support user interactions during the process of improving clustering results. We plan to develop visual analytics systems for large-scale interactive clustering by effectively leveraging the proposed methods in the human-in-the-loop analysis applications (Choo et al. 2013).

Acknowledgments The work of these authors was supported in part by NSF Grants CCF-0808863, CCF-1348152, IIS-1242304, and IIS-1231742, NIH Grant R21CA175974, and DARPA XDATA Grant FA8750-12-2-0309. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Aggarwal CC, Reddy CK (eds) (2013) Data clustering: algorithms and applications. Chapman and Hall/CRC Press, Boca Raton

Alqadah F, Bader JS, Anand R, Reddy CK (2012) Query-based biclustering using formal concept analysis. In: Proceedings of the SIAM international conference on data mining (SDM), pp 648–659

Basu S, Bilenko M, Mooney RJ (2004) A probabilistic framework for semi-supervised clustering. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 59–68

Basu S, Davidson I, Wagstaff K (2008) Constrained clustering: advances in algorithms, theory, and applications. Chapman & Hall/CRC Press, Boca Raton

Bertsekas DP (1999) Nonlinear programming, 2nd edn. Athena Scientific, Belmont
Bilenko M, Basu S, Mooney RJ (2004) Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the international conference on machine learning (ICML), pp 81–88
Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res (JMLR) 3:993–1022
Cai D, He X, Han J, Huang T (2011) Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell (TPAMI) 33(8):1548–1560
Chakrabarti D, Kumar R, Tomkins A (2006) Evolutionary clustering. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 554–560
Chen Y, Rege M, Dong M, Hua J (2008) Non-negative matrix factorization for semi-supervised data clustering. Knowl Inf Syst (KAIS) 17:355–379
Chen Y, Wang L, Dong M (2010) Non-negative matrix factorization for semi-supervised heterogeneous data coclustering. IEEE Trans Knowl Data Eng (TKDE) 22(10):1459–1474
Chi Y, Song X, Zhou D, Hino K, Tseng BL (2009) On evolutionary spectral clustering. ACM Trans Knowl Discov Data (TKDD) 3:17:1–17:30
Choo J, Lee C, Reddy CK, Park H (2013) UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans Vis Comput Graph (TVCG) 19(12):1992–2001
Guan N, Tao D, Luo Z, Yuan B (2011) Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. IEEE Trans Image Process (TIP) 20(7):2030–2048
Gupta M, Gao J, Sun Y, Han J (2012) Integrating community matching and outlier detection for mining evolutionary community outliers. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 859–867
Hoyer PO (2004) Non-negative matrix factorization with sparseness constraints. J Mach Learn Res (JMLR) 5:1457–1469
Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23(12):1495–1502
Kim J, Park H (2008) Sparse nonnegative matrix factorization for clustering. Technical report, Georgia Institute of Technology
Kim J, Park H (2011) Fast nonnegative matrix factorization: an active-set-like method and comparisons. SIAM J Sci Comput 33(6):3261–3281
Kim J, He Y, Park H (2014) Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J Glob Optim 58(2):285–319
Kuang D, Ding C, Park H (2012) Symmetric nonnegative matrix factorization for graph clustering. In: Proceedings of the SIAM international conference on data mining (SDM), pp 106–117
Kuang D, Yun S, Park H (2014) SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. J Glob Optim (to appear)
Kulis B, Basu S, Dhillon I, Mooney R (2009) Semi-supervised graph clustering: a kernel approach. Mach Learn 74(1):1–22
Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
Lee H, Yoo J, Choi S (2010) Semi-supervised nonnegative matrix factorization. IEEE Signal Process Lett 17(1):4–7
Li T, Ding C, Jordan M (2007) Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: Proceedings of the IEEE international conference on data mining (ICDM), pp 577–582
Lin CJ (2007) Projected gradient methods for nonnegative matrix factorization. Neural Comput 19(10):2756–2779
Liu H, Wu Z, Li X, Cai D, Huang T (2012) Constrained nonnegative matrix factorization for image representation. IEEE Trans Pattern Anal Mach Intell (TPAMI) 34(7):1299–1311
Liu Y, Jin R, Yang L (2006) Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Proceedings of the national conference on artificial intelligence, pp 421–426
Mimno D, McCallum A (2012) Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv:1206.3278
Shahnaz F, Berry MW, Pauca V, Plemmons RJ (2006) Document clustering using nonnegative matrix factorization. Inf Process Manag (IPM) 42(2):373–386
Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell (TPAMI) 22(8):888–905

Wang C, Yan S, Zhang L, Zhang H (2009) Non-negative semi-supervised learning. In: Proceedings of the international conference on artificial intelligence and statistics (AISTATS), pp 575–582

Xie X, Beni G (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell (TPAMI) 13(8):841–847

Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the ACM SIGIR international conference on research and development in informaion retrieval (SIGIR), pp 267–273

Zelnik-Manor L, Perona P (2004) Self-tuning spectral clustering. In: Advances in neural information processing system (NIPS), vol 17, pp 1601–1608

Zeng J, Cheung W, Li CH, Liu J (2009) Multirelational topic models. In: Proceedings of the IEEE international conference on data mining (ICDM), pp 1070–1075