Solving Consensus and Semi-supervised Clustering Problems Using Nonnegative Matrix Factorization *

Tao Li
School of CS
Florida International Univ.
Miami, FL 33199, USA
taoli@cs.fiu.edu

Chris Ding
CSE Dept.
Univ. of Texas at Arlington
Arlington, TX 76019, USA
chqding@uta.edu

Michael I. Jordan
Dept. of EECS and Dept. of Statistics
Univ. of California at Berkeley
Berkeley, CA 94720, USA
jordan@cs.berkeley.edu

Abstract

Consensus clustering and semi-supervised clustering are important extensions of the standard clustering paradigm. Consensus clustering (also known as aggregation of clustering) can improve clustering robustness, deal with distributed and heterogeneous data sources and make use of multiple clustering criteria. Semi-supervised clustering can integrate various forms of background knowledge into clustering. In this paper, we show how consensus and semi-supervised clustering can be formulated within the framework of nonnegative matrix factorization (NMF). We show that this framework yields NMF-based algorithms that are: (1) extremely simple to implement; (2) provably correct and provably convergent. We conduct a wide range of comparative experiments that demonstrate the effectiveness of this NMF-based approach.

1 Introduction

Consensus clustering and semi-supervised clustering have emerged as important elaborations of the classical clustering problem. *Consensus clustering*, also called *aggregation of clustering*, refers to the situation in which a number of different clusterings have been obtained for a particular dataset and it is desired to find a single clustering which is a good fit in some sense to the existing clusterings. Many additional problems can be reduced to the problem of consensus clustering; these include ensemble clustering, clustering of heterogeneous data sources, clustering with multiple criteria, distributed clustering, three-way clustering, and knowledge reuse [13, 11, 18, 10]. *Semi-supervised clustering* refers to the situation in which constraints are imposed on pairs of data points; in particular, there may be "must-link constraints" (two data points must be clustered into the same cluster) and "cannot-link constraints" (two data points can not be clustered into the same cluster) [19].

In this paper, we show that the consensus clustering and semi-supervised clustering problems can be usefully approached from the point of view of nonnegative matrix factorization. Nonnegative matrix factorization (NMF) refers to the problem of factorizing a given nonnegative data matrix X into two matrix factors, i.e., $X \approx AB$, while requiring A and B to be nonnegative. Originally proposed for finding parts-of-whole decompositions of images, NMF has been shown to be useful in a variety of applied settings [16, 20, 12, 2]. Algorithmic extensions of NMF have been developed to accommodate a variety of objective functions [3, 7] and a variety of data analysis problems. Based on the connection to NMF, we develop simple, provably-convergent algorithms for solving the consensus clustering and semi-supervised clustering problems. We conduct experiments on real world datasets to demonstrate the effectiveness of the new algorithms.

*Tao Li is partially supported by a IBM Faculty Research Award, NSF CAREER Award IIS-0546280 and NIH/NIGMS S06 GM008205. Chris Ding is supported in part by a University of Texas STARS Award.

2 NMF-Based Formulation of Consensus Clustering

Formally let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of n data points. Suppose we are given a set of T clusterings (or partitions) $\mathcal{P} = \{P^1, P^2, \cdots, P^T\}$ of the data points in X. Each partition $P^t, t = 1, \cdots, T$, consists of a set of clusters $C^t = \{C_1^t, C_2^t, \cdots, C_k^t\}$ where k is the number of clusters for partition P^t and $X = \bigcup_{\ell=1}^{k} C_\ell^t$. Note that the number of clusters k could be different for different clusterings.

There are several equivalent definitions of objective functions for aggregation of clustering. Following [11], we define the distance between two partitions P^1, P^2 as $d(P^1, P^2) = \sum_{i,j=1}^{n} d_{ij}(P^1, P^2)$, where the element-wise distance is defined as

$$d_{ij}(P^1, P^2) = \begin{cases} 1 & (i,j) \in C_k(P^1) \text{ and } (i,j) \notin C_k(P^2) \\ 1 & (i,j) \in C_k(P^2) \text{ and } (i,j) \notin C_k(P^1) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $(i,j) \in C_k(P^1)$ means that i and j belong to the same cluster in partition P^1 and $(i,j) \notin C_k(P^1)$ means that i and j belong to different clusters in partition P^1.

A simpler approach is to define the *connectivity matrix* as

$$M_{ij}(P^t) = \begin{cases} 1 & (i,j) \in C_k(P^t) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

We can easily see that

$$\begin{aligned} d_{ij}(P^1, P^2) &= |M_{ij}(P^1) - M_{ij}(P^2)| \\ &= [M_{ij}(P^1) - M_{ij}(P^2)]^2 \end{aligned}$$

because $|M_{ij}(P^1) - M_{ij}(P^2)| = 0$ or 1.

We look for a consensus partition (consensus clustering) P^* which is the closest to all the given partitions:

$$\min_{P^*} J = \frac{1}{T}\sum_{t=1}^{T} d(P^t, P^*) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i,j=1}^{n}[M_{ij}(P^t) - M_{ij}(P^*)]^2.$$

Let $U_{ij} = M_{ij}(P^*)$ denote the solution to this optimization problem. U is a connectivity matrix. Let the consensus (average) association between i and j be $\widetilde{M}_{ij} = \frac{1}{T}\sum_{t=1}^{T} M_{ij}(P^t)$. Define the average squared difference from the consensus association $\widetilde{M}$: $\Delta M^2 = \frac{1}{T}\sum_t \sum_{ij}[M_{ij}(P^t) - \widetilde{M}_{ij}]^2$. Clearly, the smaller ΔM^2, the closer to each other the partitions are. This quantity is a constant. We have

$$\begin{aligned} J &= \frac{1}{T}\sum_t \sum_{ij}(M_{ij}(P^t) - \widetilde{M}_{ij} + \widetilde{M}_{ij} - U_{ij})^2 \\ &= \Delta M^2 + \sum_{i,j}(\widetilde{M}_{ij} - U_{ij})^2. \end{aligned}$$

Therefore consensus clustering takes the form of the following optimization problem:

$$\min_U \sum_{i,j=1}^{n} (\widetilde{M}_{ij} - U_{ij})^2 = ||\widetilde{M} - U||^2.$$

where the matrix norm is the Frobenius norm. Therefore consensus clustering is equivalent to clustering the consensus association.

There are several formulations of consensus; for a survey see [13]. Our presentation here focuses on a simple approach. Our contribution is to show that NMF can effectively deal with the complex constraints on U_{ij} that arise in this problem.

2.1 Dealing with Constraints

Let U denote a solution of the consensus clustering problem. Being a connectivity matrix, U is characterized by a set of constraints. Consider any three nodes i, j, k. Suppose i, j belong to the same cluster: $U_{ij} = 1$. If j and k belong to the same cluster ($U_{jk} = 1$), then i and k must belong to the same cluster ($U_{ik} = 1$). On the other hand, if j and k belong to a different cluster ($U_{jk} = 0$), then i and k must belong to a different cluster ($U_{ik} = 0$). These two conditions can be expressed as

$$U_{ij} = 1,\ U_{ik} = 1,\ U_{jk} = 1 \tag{3}$$
$$U_{ij} = 1,\ U_{ik} = 0,\ U_{jk} = 0 \tag{4}$$

Now suppose that i and j belong to different clusters: $U_{ij} = 0$. We have three possibilities:

$$U_{ij} = 0,\ U_{ik} = 1,\ U_{jk} = 0 \tag{5}$$
$$U_{ij} = 0,\ U_{ik} = 0,\ U_{jk} = 1 \tag{6}$$
$$U_{ij} = 0,\ U_{ik} = 0,\ U_{jk} = 0 \tag{7}$$

These five feasibility conditions can be combined into three inequality constraints:

$$U_{ij} + U_{jk} - U_{ik} \le 1, \tag{8}$$
$$U_{ij} - U_{jk} + U_{ik} \le 1, \tag{9}$$
$$-U_{ij} + U_{jk} + U_{ik} \le 1. \tag{10}$$

There are on the order of n^3 of these inequality constraints. Solving the optimization problem satisfying these order n^3 constraints could be quite difficult.

2.2 NMF Formulation

Fortunately, these constraints can be imposed in a different way which is easy to enforce. The clustering solution can be specified by clustering indicators $H = \{0,1\}^{n\times k}$, with the constraint that in each row of H there can be only one "1" and the other entries must be zeros.

Now it is easy to show that

$$U = HH^T, \quad \text{or} \quad U_{ij} = (HH^T)_{ij}. \tag{11}$$

First, we note $(HH^T)_{ij}$ is equal to the inner product between row i of H and row j of H. Second, we consider two cases. (a) When i and j belong to the same cluster, then row i must be identical to row j; in this case $(HH^T)_{ij} = 1$. (b) When i and j belong to different clusters, the inner product between row i and row j is zero.

With $U = HH^T$, the consensus clustering problem becomes

$$\min_H ||\widetilde{M} - HH^T||^2 \tag{12}$$

where H is restricted to an indicator matrix.

Now, let us consider the relaxation of the above integer optimization. The constraint that in each row of H there is only one nonzero element can be expressed as $(H^TH)_{k\ell} = 0$ for $k \neq \ell$. Also $(H^TH)_{kk} = |C_k| = n_k$. Let

$$D = \text{diag}(H^TH) = \text{diag}(n_1, \cdots, n_k).$$

We have $H^TH = D$. Now, we can write the optimization problem as

$$\min_{H^TH=D, H\ge 0} ||\widetilde{M} - HH^T||^2 \tag{13}$$

where H is relaxed into a continuous domain.

The optimization in Eq. (13) is easier to solve than the optimization of Eq. (12). However, in Eq. (13) we need to prespecify D (the cluster sizes). But until the problem is solved we do not know D. Therefore we need to eliminate D. For this purpose, we define

$$\widetilde{H} = H(H^TH)^{-1/2},$$

Thus

$$HH^T = \widetilde{H}D\widetilde{H}^T,\ \widetilde{H}^T\widetilde{H} = H(H^TH)^{-1}H = I.$$

Therefore, the consensus clustering becomes the optimization

$$\min_{\widetilde{H}^T\widetilde{H}=I,\ \widetilde{H}, D\ge 0} ||\widetilde{M} - \widetilde{H}D\widetilde{H}^T||^2, \text{ s.t. } D \text{ diagonal.} \tag{14}$$

Now both $\widetilde{H}^T$ and D are obtained as solutions of the problem. We do not need to pre-specify the cluster sizes.

We have shown that the consensus clustering problem is equivalent to a symmetric nonnegative matrix factorization problem. In Section 3, we describe an algorithm to solve the optimization problem in Eq. (14).

2.3 Beyond Consensus Clustering

In this section we show that there are many other problems that lead to an optimization problem of the form given by Eq. (14). We consider the simplified case

$$\min_{\widetilde{H}^T\widetilde{H}=I,\ \widetilde{H}\geq 0} ||\widetilde{M}-\widetilde{H}\widetilde{H}^T||^2 \tag{15}$$

2.3.1 Kernel K-means Clustering

The average consensus similarity matrix $\widetilde{M}_{ij}$ indicates, for each pair of points, the proportion of times they are clustered together. From this point of view, $\widetilde{M}_{ij}$ measures some kind of association between i and j, and is a "similarity" between i and j.

Given a matrix W of pairwise similarities, it is known [4] that symmetric NMF

$$\min_{Q^TQ=I,Q\geq 0} ||W-QQ^T||^2. \tag{16}$$

is equivalent to kernel K-means clustering with W as the kernel: $W_{ij}=\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$, where $\mathbf{x}\rightarrow\phi(\mathbf{x})$ defines the kernel mapping.

2.3.2 Normalized Cut Spectral Clustering

Furthermore, the following symmetric NMF problem

$$\min_{Q^TQ=I,Q\geq 0} ||\widetilde{W}-QQ^T||^2, \tag{17}$$

where

$$\widetilde{W}=D^{-1/2}WD^{-1/2},\ D=\mathrm{diag}(d_1,\cdots,d_n),\ d_i=\sum_j w_{ij}$$

is equivalent to Normalized Cut spectral clustering. Thus Eq. (15) and Eq. (14) can be interpreted as clustering problems.

2.3.3 Semidefinite Programming

Note that in the quadratic form

$$||\widetilde{W}-QQ^T||^2=||\widetilde{W}||^2+Q^TQ)-2\mathrm{Tr}(Q^T\widetilde{W}Q)$$

the first two terms are constants. Thus Eq. (17) becomes

$$\max_{Q^TQ=I,Q\geq 0} \mathrm{Tr}(Q^T\widetilde{W}Q). \tag{18}$$

Xing and Jordan [21] propose to solve this problem through semidefinite programming. They write

$$\max_Q \mathrm{Tr}(Q^T\widetilde{W}Q)=\max_Q \mathrm{Tr}(\widetilde{W}QQ^T)=\max_Z \mathrm{Tr}(\widetilde{W}Z)$$

where $Z=QQ^T$ is positive semidefinite. This is a convex semidefinite programming problem and a global solution can be computed. However, once Z is obtained, there is still a challenging problem of recovering Q from Z. Clearly, this can be formulated as $\min_Q ||Z-QQ^T||^2$ which is the same as Eq. (15).

3 Algorithm for Consensus Clustering

The consensus clustering problem of Eq. (14) can be solved by reducing it to the symmetric NMF problem:

$$\min_{Q\geq 0,S\geq 0} \|W-QSQ^T\|^2,\ \ s.t.\ Q^TQ=I. \tag{19}$$

In Eq. (14) D is constrained to be a nonnegative diagonal matrix. More generally, we can relax D to be a generic symmetric nonnegative matrix.

The optimization problem in Eq. (19) can be solved using the following multiplicative update procedure:

$$Q_{jk}\leftarrow Q_{jk}\sqrt{\frac{(WQS)_{jk}}{(QQ^TWQS)_{jk}}}, \tag{20}$$

$$S_{k\ell}\leftarrow S_{k\ell}\sqrt{\frac{(Q^TWQ)_{k\ell}}{(Q^TQSQ^TQ)_{k\ell}}}. \tag{21}$$

Note that S is not restricted to being a diagonal matrix. However, if at some point during the multiplicative update procedure S becomes diagonal, it will remain that way. This feature is utilized in the algorithm. The algorithm is derived from the update algorithm for the following bi-orthogonal three-factor NMF problem, which is established to be correct and convergent in [8].

4 Semi-supervised Clustering

A key problem in semi-supervised learning is that of enforcing must-link and cannot-link constraints in the framework of K-means clustering. In many approaches the constraints are added as penalty terms in the clustering objective function, and they are iteratively enforced. In this paper, we show that the NMF perspective allows these two constraints to be enforced in a very simple and natural way within a centroid-less K-means clustering algorithm [22].

4.1 Centroid-less Constrained K-means Clustering

We again represent solutions to clustering problems using a cluster membership indicator matrix: $H=(\mathbf{h}_1,\cdots,\mathbf{h}_\kappa)$, where

$$\mathbf{h}_k=(0,\cdots,0,\overbrace{1,\cdots,1}^{n_k},0,\cdots,0)^T/n_k^{1/2} \tag{22}$$

For example, the nonzero entries of $\mathbf{h}_1$ gives the data points belonging to the first cluster. For K-means and kernel K-means the clustering objective function becomes

$$\max_{H^TH=I,\ H\geq 0} J_k=\mathrm{Tr}(H^TWH), \tag{23}$$

where $W=(w_{ij});w_{ij}=\mathbf{x}_i^T\mathbf{x}_j$ for K-means and $w_{ij}=\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ for kernel K-means .

In semi-supervised clustering, one performs clustering under two types of constraints [19]: (1) *Must-link constraints* encoded by a matrix

$$A=\{(i_1,j_1),\cdots,(i_a,j_a)\},a=|A|,$$

containing pairs of data points, where $\mathbf{x}_{i_1}, \mathbf{x}_{j_1}$ are considered similar and must be clustered into the same cluster, and (2) *Cannot-link constraints* encoded by a matrix

$$B = \{(i_1, j_1), \cdots, (i_b, j_b)\}, b = |B|,$$

where each pair of points are considered dissimilar and are not to be clustered into the same cluster.

It turns out that both must-link and cannot-link constraints can be nicely implemented in the framework of Eq. (23). A must-link pair (i_1, j_1) implies that the overlap $h_{i_1 k} h_{j_1 k} > 0$ for some k. Violation of this contraint implies that $\sum_{k=1}^{K} h_{i_1 k} h_{j_1 k} = (HH^T)_{i_1 j_1} = 0$. Therefore, we enforce the must-link constraints using the following optimization

$$\max_H \sum_{(ij)\in A} (HH^T)_{ij} = \sum_{ij} A_{ij} (HH^T)_{ij} = \mathrm{Tr} H^T A H.$$

Similarly, a cannot-link constraint for the pair (i_2, j_2) implies that $h_{i_2 k} h_{j_2 k} = 0$ for $k = 1, \cdots, K$. Thus $\sum_{k=1}^{K} h_{i_2 k} h_{j_2 k} = (HH^T)_{i_2 j_2} = 0$. Violation of this contraint implies $(HH^T)_{i_2 j_2} > 0$. Therefore we enforce the the cannot-link constraints using the optimization

$$\min_H \sum_{(ij)\in B} (HH^T)_{ij} = \sum_{ij} B_{ij} (HH^T)_{ij} = \mathrm{Tr} H^T B H.$$

Putting these conditions together, we can cast the semi-supervised clustering problem as the following optimization problem

$$\max_{H^T H = I, H \geq 0} \mathrm{Tr}[H^T W H + \alpha H^T A H - \beta H^T B H], \qquad (24)$$

where the parameter α weights the must-link constraints in A and β weights the cannot-link constraints in B.

4.2 NMF-Based Algorithm

Letting

$$W^+ = W + \alpha A \geq 0, \quad W^- = \beta B \geq 0$$

we write the semi-supervised clustering problem as follows:

$$\max_{H^T H = I, H \geq 0} \mathrm{Tr}[H^T (W^+ - W^-) H], \qquad (25)$$

Instead of solving this particular formulation, we transform to standard NMF. Note that we have the equality:

$$2\mathrm{Tr}[H^T (W^+ - W^-) H]$$
$$= ||W^+ - W^-||^2 + ||H^T H||^2 - ||(W^+ - W^-) - HH^T||^2,$$

Because the first term is a constant, and the second term is also a constant due to $H^T H = I$. Therefore, the optimization of Eq. (25) becomes

$$\min_{H^T H = I, H \geq 0} ||(W^+ - W^-) - HH^T||^2. \qquad (26)$$

This is equivalent to Eq. (25). Finally, we solve a relaxed version of Eq. (26):

$$\min_{H \geq 0} ||(W^+ - W^-) - HH^T||^2, \qquad (27)$$

where the orthogonality constraint is ignored.

4.2.1 Algorithm Description

Algorithm 2. Given an existing solution or an initial guess, we iteratively improve the solution by updating the variables with the following rule,

$$H_{ik} \leftarrow H_{ik} \sqrt{\frac{(W^+ H)_{ik}}{(W^- H)_{ik} + (HH^T H)_{ik}}}. \qquad (28)$$

4.2.2 Proof of Correctness and Convergence

We first consider the generic rectangular matrix $X = X^+ - X^-$, where $X^+ \geq 0$ and $X^- \geq 0$. We seek the factorization

$$\min_{F,G \geq 0} ||(X^+ - X^-) - FG^T||^2. \qquad (29)$$

The update rules are

$$G_{ik} \leftarrow G_{ik} \sqrt{\frac{((X^+)^T F)_{ik}}{((X^-)^T F)_{ik} + [GF^T F]_{ik}}}. \qquad (30)$$

$$F_{ik} \leftarrow F_{ik} \sqrt{\frac{(X^+ G)_{ik}}{(X^- G)_{ik} + [FG^T G]_{ik}}}. \qquad (31)$$

When X becomes symmetric, i.e., $X = X^T = W$, $F = G = H$, and both updating rules Eq. (30) and Eq. 31) become Eq. (28).

We prove the correctness and convergence of updating rule Eq. (30). Note that the proof of Eq. (31) is identical, by considering

$$\min_{F,G \geq 0} ||X^T - GF^T||^2.$$

The correctness of this algorithm can be stated as

Theorem 1 *If the solution using update rule in Eq. (30) converges, the solution satisfies the KKT optimality condition.*

Theorem 2 *The update rule Eq. (30) converges.*

A detailed proof can be found in Section 3.1 of the technical report [6].

4.3 Transitive Closure

In the remainder of this section, we discuss some theoretical aspects of our NMF-based approach for semi-supervised clustering.

Consider the case in which we enforce the constraints, i.e,

$$\alpha \gg \bar{w}; \ \beta \gg \bar{w}; \qquad (32)$$

where $\bar{w} = \sum_{ij} w_{ij}/n^2$ is the average pairwise similarity. In this case, to a first order of approximation, we can omit the first term in Eq. (24) and optimize the two constraint terms:

$$\max_H \mathrm{Tr}[\alpha H^T A H - \beta H^T B H].$$

This is further simplified into two independent optimization problems

$$\max_{H \geq 0} \mathrm{Tr}[H^T A H] \quad \text{and} \quad \min_{H \geq 0} \mathrm{Tr}[H^T B H], \qquad (33)$$

Must-link relations satisfy transitivity. Suppose $A = \{(i,j),(j,k)\}$, Then $\mathbf{x}_i$ and $\mathbf{x}_k$ should be linked. It is easy to see that this transitivity property is embedded in the solution of $\max_H \mathrm{Tr}[H^T AH]$. For example, the indicator for a specific cluster $\mathbf{h}_l$ would have nonzero entries at positions i,j,k, showing i,k are linked. Thus the solution of $\max_H \mathrm{Tr}[H^T AH]$ are the transitive closures or connected components in A.

Cannot-link relations should be consistent with must-link relations, In a cannot-link relation $B = \{(i,k)\}$, x_i, x_k can not be in the same transitive closure of A. With this requirement, the second problem in Eq. (33) is solved by simplify enforcing the constraints on B: $\mathrm{Tr}[H^T BH] = 0$.

Let $\bar{W}$ be the graph in which the C transitive closure of connected components in A are contracted into C nodes. Let $\widetilde{H}$ be cluster indicators on the nodes of contracted graph $\bar{W}$. Incorporating the solution to the first problem of Eq. (33), the solution to the whole problem is reduced to

$$\max_{\widetilde{H}^T\widetilde{H}=I, \widetilde{H}^T B\widetilde{H}=0, \widetilde{H}\geq 0} \mathrm{Tr}[\widetilde{H}^T \bar{W} \widetilde{H}]. \quad (34)$$

5 Experiments

5.1 Experiments on Consensus Clustering

The goal of this set of experiments is to evaluate the extent to which NMF-based consensus clustering can improve the robustness of traditional clustering algorithms. We compare our NMF-based consensus clustering with the results of running K-means on the original dataset, and the results of running K-means on the consensus similarity matrix. We also compare our NMF-based consensus clustering with the cluster-based similarity partitioning algorithm (CSPA), and the HyperGraph Partitioning Algorithm (HGPA) described in [17].

Dataset Description: We conduct experiments using a variety of datasets. The number of classes ranged from 2 to 20, the number of samples ranged from 47 to 4199, and the number of dimensions ranged from 4 to 1000. Further details are as follows: (i) Nine datasets (Digits389, Glass, Ionosphere, Iris, LetterIJL, Protein, Soybean, Wine, and Zoo) are from UCI data repository [9]. Digits389 is a randomly sampled subset of three classes: {3,8,9} from digits dataset. LetterIJL is a randomly sampled subset of three {I,J,L} from Letters dataset. (ii) Five datasets (CSTR, Log, Reuters, WebACE, WebKB4) are standard text datasets that are often used as benchmarks for document clustering. The documents are represented as the term vectors using the vector space model. These document datasets are pre-processed (removing the stop words and unnecessary tags and headers) using the rainbow package [15].

Analysis of Results: All the above datasets come with labels. Viewing these labels as indicative of a reasonable clustering, we use the accuracy as a performance measure [14, 5]. From Table 1, we observe that NMF-based consensus clustering improves K-means clustering on all datasets expect Reuters. Moreover, NMF-based consensus clustering achieves the best clustering performance on 9 out of 14 datasets and its performance on the remaining datasets is close to the best results. In summary, the experiments clearly demonstrate the effectiveness of NMF-based consensus clustering for improving clustering performance and robustness.

	K-means	KC	CSPA	HPGA	NMFC
CSTR	0.45	0.37	0.50	**0.62**	0.56
Digits389	0.59	0.63	**0.78**	0.38	0.73
Glass	0.38	0.45	0.43	0.40	**0.49**
Ionosphere	0.70	0.71	0.68	0.52	**0.71**
Iris	0.83	0.72	0.86	0.69	**0.89**
Protein	0.53	0.59	0.59	0.60	**0.63**
Log	0.61	**0.77**	0.47	0.43	0.71
LetterIJL	0.49	0.48	0.48	**0.53**	0.52
Reuters	**0.45**	0.44	0.43	0.44	0.43
Soybean	0.72	0.82	0.70	0.81	**0.89**
WebACE	0.41	0.35	0.40	0.42	**0.48**
WebKB4	0.60	0.56	0.61	0.62	**0.64**
Wine	0.68	0.68	0.69	0.52	**0.70**
Zoo	0.61	0.59	0.56	0.58	**0.62**

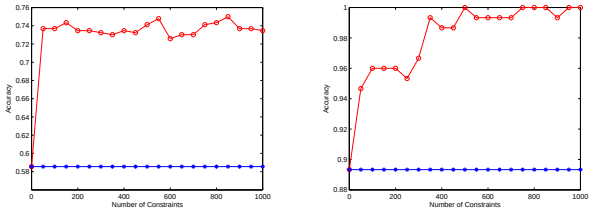
Table 1. Results on consensus clustering, as assessed by clustering accuracy. The results are obtained by averaging over five trials. KC represents the results of applying K-means to a consensus similarity matrix, and NMFC represents the NMF-based consensus clustering.

5.2 Experiments on Semi-supervised Clustering

In this section, we report the results of experiments conducted to investigate the effectiveness of NMF-based semi-supervised clustering. In our experiments, we set α, the weight for must-link constraints, to be 2, and β, the weight of cannot-link constraints, to be 1. This choice of weights implies that the cannot-link constraints are not as vigorously enforced as the must-link constraints.

We use the accuracy measure to evaluate the performance of semi-supervised clustering. Note that this measure is different from the F-measure used in previous studies (e.g., [1]). Since our goal is to discover the one-to-one relationship between generated clusters and underlying classes, and to measure the extent to which each cluster contains data points from the corresponding class, we feel accuracy is an appropriate performance measure. Figure 1 plots the clustering accuracy as a function of the number of constraints on two datasets, e.g., iris and digits. We observe that the enhancement obtained by the semi-supervised clustering is generally greater as the number of constraints increases. Similar behaviors can also be observed in other datasets. Due to space limits, we only include the curves for the above three datasets.

We also perform experiments on the following six datasets and compare the results of our NMF-based algorithm with those obtained by MPCKmeans algorithm. MPCKmeans clustering incorporates both seeding and metric learning in a uni-



(a) Digits389 Dataset (b) Iris Dataset

Figure 1. Results on semi-supervised clustering using our NMF based algorithm. Accuracy as a function of numbers of constraints.

fied framework and performs distance-metric training with each clustering iteration using the constraints [1]. Table 2 presents the results of this comparison. We observe that NMF-based semi-supervised clustering improves clustering performance. Also, though the differences are small, NMF-based semi-supervised clustering outperforms MPCKmeans on 4 out of 6 datasets.

	K-means	MPCKmeans	NMFS
Digits389	0.5855	**0.7400**	0.7346
Glass	0.3832	**0.4752**	0.4673
Iris	0.8263	0.9400	**0.9600**
Protein	0.5259	0.5517	**0.5603**
LetterIJL	0.4890	0.5178	**0.5286**
Soybean	0.7830	0.8298	**0.8936**

Table 2. Results on semi-supervised clustering, as assessed by clustering accuracy. 200 constraints are generated for each dataset. The results are obtained by averaging over five trials. NMFS represents the NMF-based semi-supervised clustering algorithm.

6 Conclusions

We have shown that consensus clustering and semi-supervised clustering can be formulated within the framework of nonnegative matrix factorization. This yields simple iterative updating algorithms for solving these problems. We demonstrated the effectiveness of the NMF-based approach in a variety of comparative experiments. Our work has expanded the scope of NMF applications in the clustering domain and has highlighted the wide applicability of this class of learning algorithms.

References

[1] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, 2004.

[2] J.-P. Brunet, P. Tamayo, T. Golub, and J. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Nat'l Academy of Sciences USA*, 102(12):4164–4169, 2004.

[3] I. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *NIPS 17*, 2005.

[4] C. Ding, X. He, and H. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. *Proc. SIAM Data Mining Conf*, 2005.

[5] C. Ding, X. He, H. Zha, and H. D. Simon. Adaptive dimension reduction for clustering high dimensional data. In *ICDM*, 2002.

[6] C. Ding, T. Li, and M. Jordan. Convex and semi-nonnegative matrix factorizations for clustering and low-dimension representation. Technical Report LBNL-60428, Lawrence Berkeley National Laboratory, University of California, Berkeley, 2006.

[7] C. Ding, T. Li, and W. Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. *Proc. National Conf. Artificial Intelligence*, 2006.

[8] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *SIGKDD*, pages 126–135, 2006.

[9] C. B. D.J. Newman, S. Hettich and C. Merz. UCI repository of machine learning databases, 1998.

[10] X. Z. Fern and C. E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *ICML*, 2004.

[11] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. In *ICDE*, pages 341–352, 2005.

[12] S. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *CVPR*, pages 207–212, 2001.

[13] T. Li, M. Ogihara, and S. Ma. On combining multiple clusterings. In *CIKM*, pages 294–303, 2004.

[14] T. Li, M. Ogihara, and S. Zhu. Integrating features from different sources for music information retrieval. In *ICDM*, pages 372–381, 2006.

[15] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[16] P. Paatero and U. Tapper. Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[17] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *JMLR*, 3:583–617, December 2002.

[18] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *JMLR*, 3:583–617, March 2003.

[19] K. Wagsta, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *ICML*, pages 577–584, 2001.

[20] Y.-L. Xie, P. Hopke, and P. Paatero. Positive matrix factorization applied to a curve resolution problem. *Journal of Chemometrics*, 12(6):357–364, 1999.

[21] E. Xing and M. Jordan. On semidefinite relaxation for normalized k-cut and connections to spectral clustering. *University of California Berkeley Tech Report CSD-03-1265*, 2003.

[22] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for K-means clustering. *NIPS 14)*, 2002.