BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Non-negative matrix factorization by maximizing correntropy for cancer clustering

Jim Jing-Yan Wang[1], Xiaolei Wang[1] and Xin Gao[1,2*]

Abstract

Background: Non-negative matrix factorization (NMF) has been shown to be a powerful tool for clustering gene expression data, which are widely used to classify cancers. NMF aims to find two non-negative matrices whose product closely approximates the original matrix. Traditional NMF methods minimize either the l_2 norm or the Kullback-Leibler distance between the product of the two matrices and the original matrix. Correntropy was recently shown to be an effective similarity measurement due to its stability to outliers or noise.

Results: We propose a maximum correntropy criterion (MCC)-based NMF method (NMF-MCC) for gene expression data-based cancer clustering. Instead of minimizing the l_2 norm or the Kullback-Leibler distance, NMF-MCC maximizes the correntropy between the product of the two matrices and the original matrix. The optimization problem can be solved by an expectation conditional maximization algorithm.

Conclusions: Extensive experiments on six cancer benchmark sets demonstrate that the proposed method is significantly more accurate than the state-of-the-art methods in cancer clustering.

Background

Because cancer has been a leading cause of death in the world for several decades, the classification of cancers is becoming more and more important to cancer treatment and prognosis [1,2]. With advances in DNA microarray technology, it is now possible to monitor the expression levels of a large number of genes at the same time. There have been a variety of studies on analyzing DNA microarray data for cancer class discovery [3-5]. Such methods are demonstrated to outperform the traditional, morphological appearance-based cancer classification methods. In such studies, different cancer classes are discriminated by their corresponding gene expression profiles [1].

Several clustering algorithms have been used to identify groups of similar expressed genes. Non-negative matrix factorization (NMF) was recently introduced to analyze gene expression data and this method demonstrated superior performance in terms of both accuracy and stability [6-8]. Gao and Church [3] reported an effective unsupervised method for cancer clustering with gene expression profiles via sparse NMF (SNMF). Carmona et al. [9] presented a methodology that was able to cluster closely related genes and conditions in sub-portions of the data based on non-smooth non-negative matrix factorization (nsNMF), which was able to identify localized patterns in large datasets. Zheng et al. [5,7] applied penalized matrix decomposition (PMD) to extract meta-samples from gene expression data, which could captured the inherent structures of samples that belonged to the same class.

NMF approximates a given gene data matrix, X, as a product of two low-rank nonnegative matrices, H and W, as $X \approx HW$. This is usually formulated as an optimization problem, where the objective function is to minimize either the l_2 norm or the Kullback-Leibler (KL) distance [10] between X and HW. Most of the improved NMF algorithms are also based on the minimization of these two distances while adding the sparseness term [3], the graph regularization term [11], etc. Sandler and Lindenbaum [12] argued that measuring the dissimilarity of W and HW by either the l_2 norm or the KL distance, even with additional bias terms, was inappropriate in computer vision applications due to the nature of errors in images. Sandler

*Correspondence: xin.gao@kaust.edu.sa
[1]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[2]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

and Lindenbaum [12] proposed a novel NMF with earth mover's distance (EMD) metric by minimizing the EMD error between X and HW. The proposed NMF-EMD algorithm demonstrated significantly improved performance in two challenging computer vision tasks, i.e., texture classification and face recognition. Liu et al. [4] tested a family of NMF algorithms using α-divergence with different α values as dissimilarities between X and HW for clustering cancer gene expression data.

It is widely acknowledged that DNA microarry data contain many types of noise, especially experimental noise. Recently, correntropy was shown to be an effective similarity measurement in information theory due to its stability to outliers or noise [13]. However, it has not been used in the analysis of microarray data. In this paper, we propose a novel form of NMF that maximizes the correntropy. We introduce a new NMF algorithm with a maximum correntropy criterion (MCC) [13] for the gene expression data-based cancer clustering problem. We call it NMF-MCC. The goal of NMF-MCC is to find a meta-sample matrix, H, and a coding matrix, W, such that the gene expression data matrix, X, is as correlative to the product of H and W as possible under MCC.

Related works

He et al. [13] recently developed a face recognition algorithm, correntropy-based sparse representation (CESR), based on MCC. CESR tries to find a group of sparse combination coefficients to maximize the correntropy between the facial image vector and the linear combination of faces in the database. He et al. [13] demonstrated that CESR was much more effective in dealing with the occlusion and corruption problems of face recognition than the state-of-the-art methods. However, CESR learns only the combination coefficients while the basis faces (the faces in the database) are fixed. Comparing to CESR, NMF-MCC can learn both the combination coefficients and the basis vectors jointly, which allows the algorithm to obtain more basis vectors for better representation of the data points. Zafeiriou and Petrou [14] addressed the problem of NMF with kernel functions instead of inner products and proposed the projected gradient kernel nonnegative matrix factorization (PGK-NMF) algorithm. Both NMF-MCC and PGK-NMF employ kernel functions to map the linear data space to a non-linear space. However, as we show later, NMF-MCC computes different kernels for different features, while PGK-NMF computes a single kernel for the whole feature vector. Thus, NMF-MCC allows the algorithm to assign different weights to different features and emphasizes the discriminant features with high weights, thus achieving feature selection. In contrast, like most kernel based methods, PGK-NMF simply replaces the inner product by the kernel-function and treats the features equally, thus there is no feature selection function.

Methods

In this section, we first briefly introduce the traditional NMF method. We then propose our novel NMF-MCC algorithm by maximizing the correntropy in NMF. We further propose a expectation conditional maximization-based approach to solve the optimization problem.

Nonnegative matrix factorization

NMF is a matrix factorization algorithm that focuses on the analysis of data matrices whose elements are nonnegative. Consider a gene expression dataset that consists of D genes in N samples. We denote it by a matrix $X = [x_1, \cdots, x_N] \in \Re^{D\times N}$ of size $D \times N$, and each column of X is a sample vector containing D genes. NMF aims to find two non-negative matrices, $H = [h_{dk}] \in \Re^{D\times K}$ and $W = [w_{kn}] \in \Re^{K\times N}$, whose product closely approximates the original matrix X:

$$X \approx HW. \tag{1}$$

Matrix H is of size $D \times K$, with each of the K columns defining a meta-sample and each entry, h_{dk}, in H representing the expression level of gene d over meta-sample k. Matrix W is of size $K \times N$, with each of the n columns representing the meta-sample expression pattern of the corresponding sample, and each entry, w_{kn}, representing the coefficient of meta-sample k over sample n. Figure 1 shows an example of the factorization of a gene expression matrix X with $D = 2308$ genes and $N = 83$ samples as the product of the meta-sample matrix H with $K = 4$ meta-samples and the coding matrix W.

The factorization is quantified by an objective function that minimizes some distance measure, such as:

- **l_2 norm distance**: One simple measure is the square of the l_2 norm distance (also known as the Frobenius norm or the Euclidean distance) between two matrices, which is defined as:

$$F^{l_2} = \sum_{d=1}^{D}\sum_{n=1}^{N}\left(X_{dn} - \sum_{k=1}^{K} H_{dk}W_{kn}\right)^2. \tag{2}$$

- **Kullback - Leibler (KL) divergence**: The second one is the divergence between two matrices [10], which is defined as:

$$F^{KL} = \sum_{d=1}^{D}\sum_{n=1}^{N}\left(X_{dn} ln \frac{X_{dn}}{(HW)_{dn}} - X_{dn} + (HW)_{nd}\right). \tag{3}$$

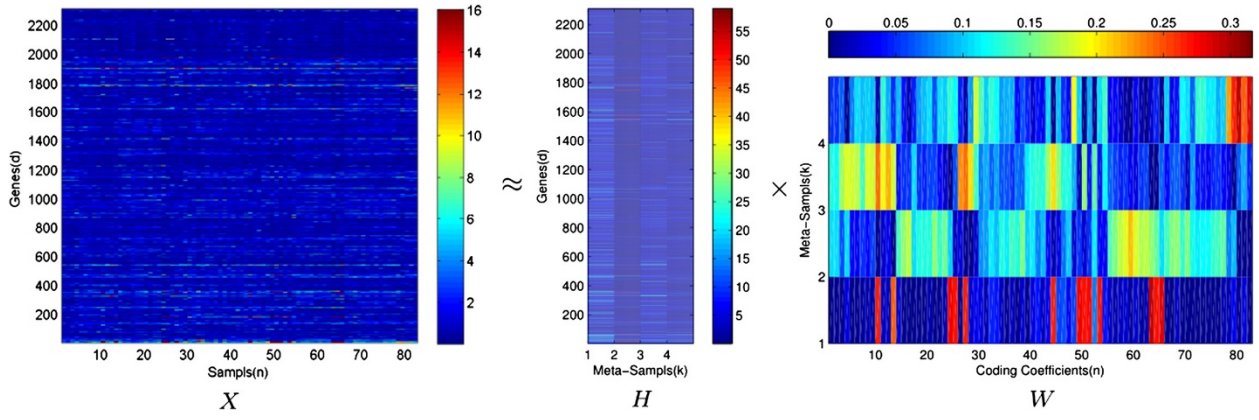

Figure 1 The l_2 norm distance-based non-negative matrix factorization on the SRBCT dataset [29]. The gene expression data matrix, X, is factorized as the product of the meta-sample matrix, H, and the coding matrix, W.

Maximum correntropy criterion for NMF

Another thing that has to be changed is that the definition of correntropy is not subject to the kernel being Gaussian as they seem to imply through the text, so for instance when they define they can say E(k(x-y)) and one of the common choices of k is the Gaussian kernel giving....

Correntropy is a nonlinear similarity measure between two random variables, x and y [13,15,16], defined as

$$V_\sigma(x,y) = E[\,k_\sigma(x-y)]\,, \tag{4}$$

where k_σ is a kernel that satisfies the Mercer theory and $E[\,\cdot]$ is the expectation. One of the common choices of k_σ is the Gaussian kernel given as $k_\sigma(x-y) = exp(-\frac{(x-y)^2}{2\sigma^2})$.

In practice, the joint probability density function of x and y is unknown and only a finite amount of data $\{(x_i, y_i)\}, i = 1, \cdots, I$ is available. Therefore, the sample correntropy is estimated by

$$\widehat{V}_\sigma(x,y) = \frac{1}{I}\sum_{i=1}^{I} k_\sigma(x_i - y_i), \tag{5}$$

Based on Eq. (5), a general similarity measurement between any two discrete gene expression vectors was proposed [17]. They introduced the correntropy induced metric (CIM) for any two gene sample vectors $x = [x_1, \cdots, x_D]^\top$ and $y = [y_1, \cdots, y_D]^\top$, as:

$$\begin{aligned} CIM(x,y) &= \left(k_\sigma(0) + \frac{1}{D}\sum_{d=1}^{D} k_\sigma(x_d - y_d)\right)^{\frac{1}{2}} \\ &= \left(k_\sigma(0) + \frac{1}{D}\sum_{d=1}^{D} k_\sigma(e_d)\right)^{\frac{1}{2}}, \end{aligned} \tag{6}$$

where $e_d = x_d - y_d$ is defined as the error. For adaptive systems, we can define the maximum correntropy criterion (MCC) [18] as

$$\begin{aligned} &\max_{\Theta} \sum_{d=1}^{D} k_\sigma(x_d - y_d), \\ &k_\sigma(x_d - y_d) = exp\left[-\frac{(x_d - y_d)^2}{2\sigma^2}\right] \end{aligned} \tag{7}$$

where Θ is a parameter to be specified later. We must notice the difference between MCC and common kernel criterion used in [14]. The Gaussian kernel function of vectors x and y is defined as

$$\begin{aligned} k_\sigma(x-y) &= exp\left(-\frac{||x-y||^2}{2\sigma^2}\right) \\ &= exp\left[-\frac{\sum_{d=1}^{D}(x_d - y_d)^2}{2\sigma^2}\right]. \end{aligned} \tag{8}$$

We can see that the kernel is applied to the entire feature vector, x, and each feature $x_d, d = 1 \cdots, D$ is treated equally with the same kernel parameter. However, in (7), kernel functions are applied to different functions. This can allow the algorithm to learn different kernel parameters as we will introduce later. In this way, we can assign different weights to different features and thus implement feature selection.

Our goal is to find a meta-sample matrix, H, and a coding matrix, W, such that HW is as correlative to X as possible under MCC as described in Eq. (7). To extend MCC from vector space R^D to matrix space $R^{D\times N}$, we replace $e_d = (x_d - y_d)$ with the l_2 norm distance between the samples of X and $Y = HW$ as $e_d = \sqrt{\sum_{n=1}^{N}(x_{dn} - y_{dn})^2}$, where y_{dn} is the (d,n)-th item of Y,

and $y_{dn} = \sum_{k=1}^{K} h_{dk} w_{kn}$. Moreover, the factorization system parameter should be set to $\Theta = (H, W)$ under the framework of NMF-MCC. By substituting newly defined e_d and Θ to (7), we can formulate the problem of NMF-MCC as the following optimization problem:

$$\begin{aligned} &\max_{H,W} F(H, W) \\ &s.t.\ H \geq 0,\ W \geq 0. \\ &F(H, W) = \sum_{d=1}^{D} k_\sigma\ (e_d) \\ &= \sum_{d=1}^{D} k_\sigma \left(\sqrt{\sum_{n=1}^{N} (x_{dn} - \sum_{k=1}^{K} h_{dk} w_{kn})^2} \right) \\ &= \sum_{d=1}^{D} exp \left(- \frac{\sum_{n=1}^{N} (x_{dn} - \sum_{k=1}^{K} h_{dk} w_{kn})^2}{2\sigma^2} \right). \end{aligned} \tag{9}$$

We should notice the significant difference between NMF-MCC and CESR. As a supervised learning algorithm, the CESR represents a test data point, x_t, as a linear combination of all the the training data points as $x_t \approx \sum_{n=1}^{N} x_n w_{nt} = X w_t$ and $w_t = [\, w_{1t}, \cdots, w_{Nt}]^\top$ is the combination coefficient vector. CESR aims to find the optimal w_t to maximize the correntropy between x_t and Xw_t. Similarly, NMF-MCC also tries to represent a data point x_n as a linear combination of some basis vectors as $x_n \approx \sum_{k=1}^{K} h_k w_{kn} = X w_n$ and $w_n = [\, w_{1n}, \cdots, w_{Kn}]^\top$ is the combination coefficient vector. Differently from CESR, NMF-MCC aims to find not only the optimal w_n but also the basis vectors in H to maximize the correntropy between x_n and Hw_n, $n = 1, \cdots, N$. The internal difference between NMF-MCC and CESR lies in whether to learn basis vectors or not.

In order to solve the optimization problem, we recognize that the expectation conditional maximization (ECM) method [19] can be applied. Based on the theory of convex conjugate functions [20], we can derive the following proposition that forms the basis to solve the optimization problem in (9):

Proposition 1. *There exists a convex conjugate function of $g(z, \sigma)$ such that*

$$g(z, \sigma) = \sup_{\varrho \in \Re^-} \left(\varrho \frac{||z||^2}{\sigma^2} - \varphi(\varrho) \right) \tag{10}$$

and for a fixed z, the supremum is reached at $\varrho = -g(z, \sigma)$.

By substituting Eq. (10) into (9), we have the augmented objective function in an enlarged parameter space

$$\begin{aligned} &\max_{H,W,\boldsymbol{\rho}} \widehat{F}(H, W, \boldsymbol{\rho}) \\ &s.t.\ H \geq 0,\ W \geq 0. \\ &\widehat{F}(H, W, \boldsymbol{\rho}) = \sum_{d=1}^{D} \left(\rho_d \sum_{n=1}^{N} (x_{dn} - \sum_{k=1}^{K} h_{dk} w_{kn})^2 - \varphi(\rho_d) \right), \end{aligned} \tag{11}$$

where superscript φ is the convex conjugate function φ of $g(z)$ defined in Proposition 1, and $\boldsymbol{\rho} = [\, \rho_1, \cdots, \rho_D]^\top$ are the auxiliary variables.

According to Proposition 1, for fixed H and W, the following equation holds:

$$F(H, W) = \max_{\boldsymbol{\rho}} \widehat{F}(H, W, \boldsymbol{\rho}). \tag{12}$$

It follows that

$$\begin{aligned} \max_{H,W} F(H, W) &= \max_{H,W} \left[\max_{\boldsymbol{\rho}} \widehat{F}(H, W, \boldsymbol{\rho}) \right] \\ &= \max_{H,W,\boldsymbol{\rho}} \widehat{F}(H, W, \boldsymbol{\rho}). \end{aligned} \tag{13}$$

That is, maximizing $F(H, W)$ is equivalent to maximizing the augmented function $\widehat{F}(H, W, \boldsymbol{\rho})$.

The NMF-MCC Algorithm

The traditional NMF can be solved by the expectation-maximization (EM) algorithm [21]. However, in the case of MCC-based NMF, EM must be replaced by ECM because there is more than one parameter. Figure 2 shows the outline of ECM, which is described in more detail below.

1. **E-Step**: Compute $\boldsymbol{\rho}$ given the current estimations of the meta-sample matrix H and the coding matrix W as:

$$\rho_d^t = -g \left(\sqrt{\sum_{n=1}^{N} \left(x_{dn} - \sum_{k=1}^{K} h_{dk}^t w_{kn}^t \right)^2}, \sigma^t \right), \tag{14}$$

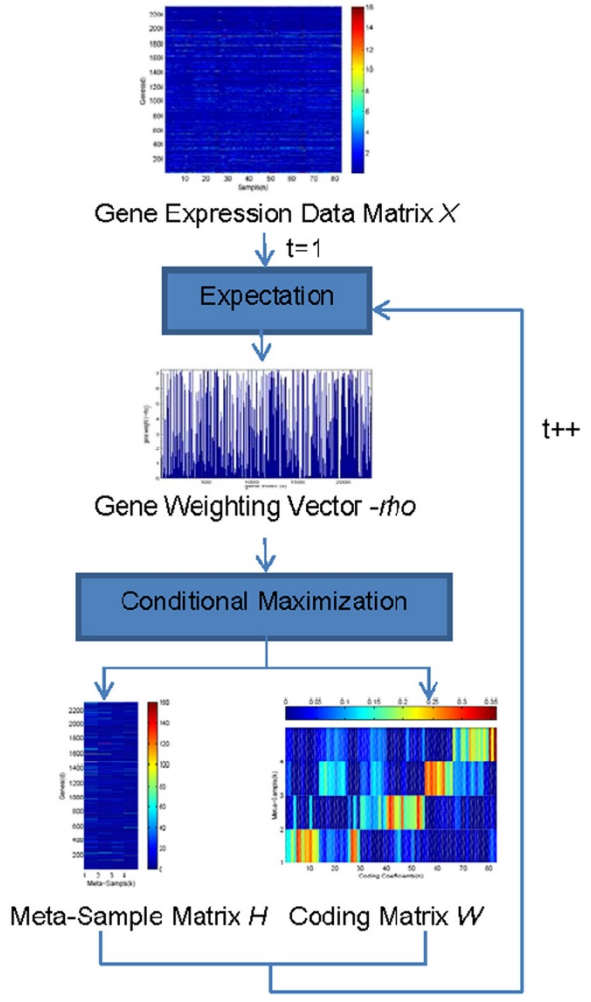

Figure 2 Outline of the ECM-based NMF-MCC algorithm.

where t means the t-th iteration. In this study, the kernel size (bandwidth) σ^{2^t} is computed by

$$\sigma^{2^t} = \frac{\theta}{2D}\sum_{d=1}^{D}\sum_{n=1}^{N}\left(x_{dn} - \sum_{k=1}^{K} h_{dk}^t w_{kn}^t\right)^2, \quad (15)$$

where θ is a parameter to control the sparseness of ρ_d^t.

2. **CM-steps**: In the CM-step, given ρ_d^t, we try to optimize the following function respect to H and W:

$$\begin{aligned}(H^{t+1}, W^{t+1}) &= \underset{H,W}{argmax}\sum_{d=1}^{D}\left(\rho_d^t\sum_{n=1}^{N}\left(x_{dn} - \sum_{k=1}^{K} h_{dk}w_{kn}\right)^2\right)\\ &= \underset{H,W}{argmax}\ Trac\left[(X-HW)^\top diag(\boldsymbol{\rho}^t)(X-HW)\right]\\ &s.t.\ H \geq 0,\ W \geq 0,\end{aligned} \quad (16)$$

where $diag(\cdot)$ is an operator that converts the vector $\boldsymbol{\rho}$ to a diagonal matrix.
By introducing a dual objective function,

$$\begin{aligned}\mathcal{O}(H,W) =& Trac\left[(X-HW)^\top diag(-\boldsymbol{\rho}^t)(X-HW)\right]\\ =& Trac\left[X^\top diag(-\boldsymbol{\rho}^t)X\right] - 2Trac\left[X^\top diag(-\boldsymbol{\rho}^t)HW\right]\\ &+ Trac\left[W^\top H^\top diag(-\boldsymbol{\rho}^t)HW\right],\end{aligned} \quad (17)$$

the optimal problem in (16) can be reformulated as the following dual problem:

$$\begin{aligned}(H^{t+1}, W^{t+1}) =& \underset{H,W}{argmin}\ \mathcal{O}(H,W)\\ &s.t.\ H \geq 0,\ W \geq 0.\end{aligned} \quad (18)$$

Let ϕ_{dk} and ψ_{kn} be the Lagrange multiplier for constraints $h_{dk} \geq 0$ and $w_{kn} \geq 0$, respectively, and $\Phi = [\phi_{dk}]$ and $\Psi = [\psi_{kn}]$. The Lagrange $\mathcal{L}$ is

$$\begin{aligned}\mathcal{L} =& Trac\left[X^\top diag(-\boldsymbol{\rho}^t)X\right] - 2Trac\left[X^\top diag(-\boldsymbol{\rho}^t)HW\right]\\ &+ Trac\left[W^\top H^\top diag(-\boldsymbol{\rho}^t)HW\right] + Trac\left[\Phi H^\top\right]\\ &+ Trac\left[\Psi W^\top\right].\end{aligned} \quad (19)$$

The partial derivatives of $\mathcal{L}$ with respect to H and W are

$$\frac{\partial \mathcal{L}}{\partial H} = -2diag(-\boldsymbol{\rho}^t)XW^\top + 2diag(-\boldsymbol{\rho}^t)HWW^\top + \Phi \quad (20)$$

and

$$\frac{\partial \mathcal{L}}{\partial W} = -2H^\top diag(-\boldsymbol{\rho}^t)X + 2H^\top diag(-\boldsymbol{\rho}^t)HW + \Psi \quad (21)$$

Using the Karush-Kuhn-Tucker optimal conditions, i.e., $\phi_{dk}h_{dk} = 0$ and $\psi_{kn}w_{kn} = 0$, we get the following equations for h_{dk} and w_{kn}:

$$\begin{aligned}&-2(diag(-\boldsymbol{\rho}^t)XW^\top)_{dk}h_{dk}\\ &+2(diag(-\boldsymbol{\rho}^t)HWW^\top)_{dk}h_{dk} = 0\end{aligned} \quad (22)$$

and

$$\begin{aligned}&-2(H^\top diag(-\boldsymbol{\rho}^t)X)_{kn}w_{kn}\\ &+2(H^\top diag(-\boldsymbol{\rho}^t)HW)_{kn}w_{kn} = 0\end{aligned} \quad (23)$$

These equations lead to the following updating rules to maximize the expectation in (13).

- The meta-sample matrix H, conditioned on the coding matrix W:

$$h_{dk}^{t+1} \longleftarrow h_{dk}^{t} \frac{(diag(-\boldsymbol{\rho}^t)XW^{t\top})_{dk}}{(diag(-\boldsymbol{\rho}^t)H^tW^tW^{t\top})_{dk}} \tag{24}$$

- The coding matrix W conditioned on the newly estimated meta-sample matrix H^{t+1}:

$$w_{kn}^{t+1} \longleftarrow w_{kn}^{t} \frac{(H^{t+1\top}diag(-\boldsymbol{\rho}^t)X)_{kn}}{(H^{t+1\top}diag(-\boldsymbol{\rho}^t)H^{t+1}W^t)_{kn}} \tag{25}$$

We should note that if we exchange the numerator and denominator in (24) and (25), new update formulas will be yield. The new update rules are dual for (24) and (25), and our experimental results show that the dual update rules achieve similar clustering performances as (24) and (25).

Algorithm 1 summarizes the optimization procedure.

Algorithm 1 NMF-MCC Algorithm.

Require: Input gene expression data matrix X;
Require: Initial meta-sample gene matrix H^1 and coding matrix W^1;
for $t = 1, \cdots, T$ **do**
  Update the auxiliary variables $\boldsymbol{\rho}^t$ as in (14);
  Update the meta-sample matrix H^{t+1} as in (24);
  Update the coding matrix W^{t+1} as in (25);
end for
Output $H = H^{T+1}$ and $W = W^{T+1}$.

Proof of convergence

In this section, we will prove that the objective function in (16) is nonincreasing under the updating rules in (24) and (25).

Theorem 1. The objective function in (16) is nonincreasing under the update rules (24) and (25).

To prove the above theorem, we first define an auxiliary function.

Definition 1. $G(w, w')$ is an auxiliary function for $F(w)$ if the conditions

$$G(w, w') \geq F(w),\ G(w, w) = F(w) \tag{26}$$

are satisfied.

The auxiliary function is quite useful because of the following lemma:

Lemma 1. If G is an auxiliary function of F, then F is nonincreasing under the update

$$w^{t+1} = \underset{w}{argmin} G(w, w^t). \tag{27}$$

We refer the readers to [22] for the proof of this lemma. Now, we show that the updating rule of (25) is exactly the update in (27) with a proper auxiliary function. We denote the objective function in (16) as O:

$$\begin{aligned} O &= \sum_{d=1}^{D}\left(\rho_d \sum_{n=1}^{N}(x_{dn} - \sum_{k=1}^{K} h_{dk}w_{kn})^2\right) \\ &= Trac\left[(X - HW)^\top diag(\boldsymbol{\rho}^t)(X - HW)\right]. \end{aligned} \tag{28}$$

Considering any element, w_{kn}, in W, we use F_{kn} to denote the part of the objective function in (16) that is relevant only to w_{kn}. It is easy to check that

$$\begin{aligned} F'_{kn} &= \left(\frac{\partial O}{\partial W}\right)_{kn} = \Big(-2H^\top diag(-\boldsymbol{\rho}^t)X \\ &\qquad +2H^\top diag(-\boldsymbol{\rho}^t)HW\Big)_{kn} \\ F''_{kn} &= \left(\frac{\partial^2 O}{\partial^2 W}\right)_{kn} = 2\left(H^\top diag(-\boldsymbol{\rho}^t)H\right)_{kk} \end{aligned} \tag{29}$$

Since the updating rule is essentially based on elements, it is sufficient to show that each F_{kn} is nonincreasing under the update step of (25).

Table 1 Summary of the six cancer gene expression datasets used to test the NMF-MCC algorithm

Dataset name	Diagnostic task	Samples (N)	Genes (D)	Cancer Classes (K)	Ref
Leukemia	Acute myelogenous leukemia	72	5327	3	[25]
Brain Tumor	5 human brain tumor types	90	5920	5	[26]
Lung Cancer	4 lung cancer types and normal tissues	203	12600	5	[27]
9 Tumors	9 various human tumor types	60	5726	9	[28]
SRBCT	Small, round blue cell tumors	83	2308	4	[29]
DLBCL	Diffuse large B-cell lymphomas	77	5469	2	[24]

Lemma 2. Function

$$G(w, w_{kn}^t) = F_{kn}^t(w_{kn}^t) + F'_{kn}(w_{kn}^t)(w - w_{kn}^t) + \frac{(H^\top diag(-\boldsymbol{\rho}^t)HW)_{kn}}{w_{kn}^t}(w - w_{kn}^t)^2 \quad (30)$$

is an auxiliary function for F_{kn}, which is relevant only to w_{kn}.

Proof. Since $G(w, w) = F_{kn}(w)$ is obvious, we only need to show that $G(w, w_{kn}^t) \geq F_{kn}(w)$. To do this, we compare the Taylor series expansion of $F_{kn}(w)$,

$$\begin{aligned} F_{kn}(w) &= F_{kn}(w_{kn}^t) + F'_{kn}(w_{kn}^t)(w - w_{kn}^t) \\ &\quad + \frac{1}{2}F''_{kn}(w_{kn}^t)(w - w_{kn}^t)^2 \\ &= F_{kn}(w_{kn}^t) + F'_{kn}(w_{kn}^t)(w - w_{kn}^t) \\ &\quad + \left(H^\top diag(-\boldsymbol{\rho}^t)H\right)_{kk}(w - w_{kn}^t)^2 \end{aligned} \quad (31)$$

with (30) to find that $G(w, w_{kn}^t) \geq F_{kn}(w)$ is equivalent to

$$\begin{aligned} \frac{(H^\top diag(-\boldsymbol{\rho}^t)HW)_{kn}}{w_{kn}^t} &\geq \left(H^\top diag(-\boldsymbol{\rho}^t)H\right)_{kk} \\ (H^\top diag(-\boldsymbol{\rho}^t)HW)_{kn} &\geq \left(H^\top diag(-\boldsymbol{\rho}^t)H\right)_{kk} w_{kn}^t \end{aligned} \quad (32)$$

We have

$$\begin{aligned} (H^\top diag(-\boldsymbol{\rho}^t)HW)_{kn} &= \sum_{l=1}^{K}(H^\top diag(-\boldsymbol{\rho}^t)H)_{kl}w_{ln}w^t \\ &\geq \left(H^\top diag(-\boldsymbol{\rho}^t)H\right)_{kk} w_{kn}^t. \end{aligned} \quad (33)$$

Thus, (32) holds and $G(w, w_{kn}^t) \geq F_{kn}(w)$. □

We can now demonstrate the convergence of **Theorem 1**.

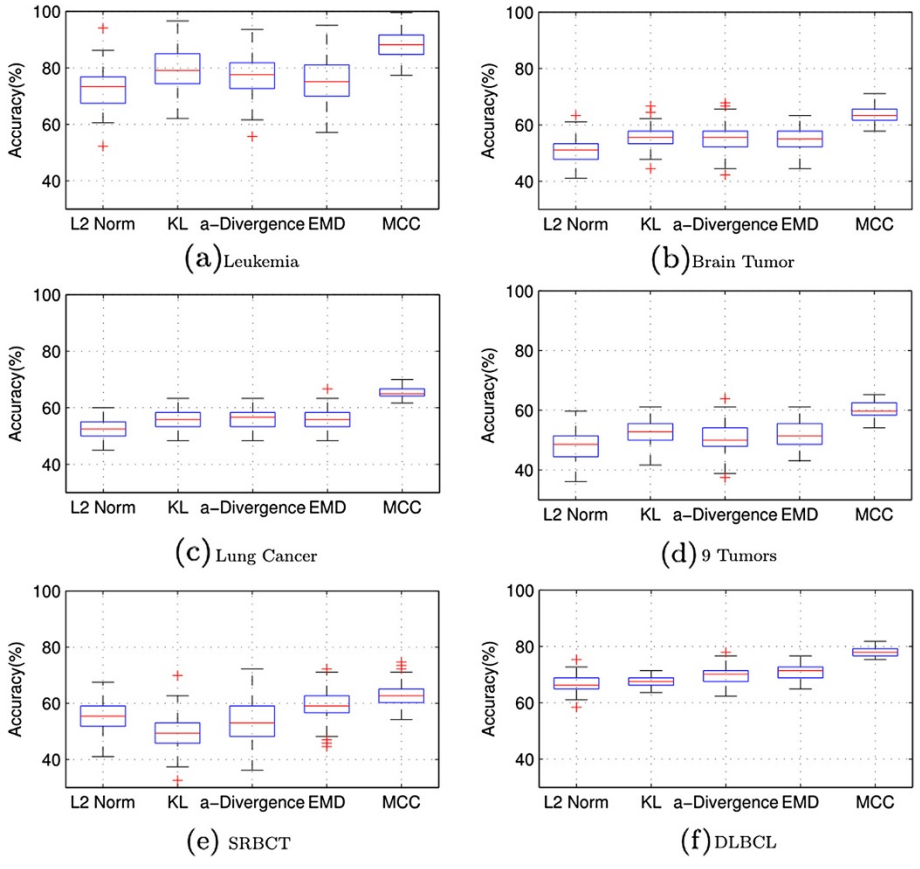

Figure 3 The boxplots of the clustering accuracies for NMF with different loss functions over 100 runs on the six gene expression datasets: (a) Leukemia, (b) Brain Tumor, (c) Lung Cancer, (d) 9 Tumors, (e) SRBCT, (f) DLBCL.

Proof of Theorem 1. Replacing $G(w, w^t)$ in (27) by (30) results in the update rule

$$w_{kn}^{t+1} = w_{kn}^t - w_{kn}^t \frac{F'_{kn}(w_{kn}^t)}{2(H^\top diag(-\boldsymbol{\rho})HW^t)_{kn}} = w_{kn}^t \frac{(H^\top diag(-\boldsymbol{\rho})X)_{kn}}{(H^\top diag(-\boldsymbol{\rho})HW^t)_{kn}}. \quad (34)$$

Since (30) is an auxiliary function, F_{kn} is nonincreasing under this update rule as in (25).

Similarly, we can also show that O is nonincreasing under the updating steps in (24).

Experiments

Datasets

To test the proposed algorithm, we carry out extensive experiments on six cancer-related gene expression datasets. The six datasets consist of five multi-class sets as used in [4,23] and one binary class set [24]. The descriptions of the six datasets are summarized in Table 1. In these datasets, besides the gene expression data samples, the labels are also given. They were obtained from the diagnosis results and reported in different studies [23].

Performance metric

The proposed NMF-MCC algorithm will be used to represent gene expression data for k-means clustering. The clustering results are evaluated by comparing the obtained label of each sample with the label provided by the dataset. The clustering accuracy is used to measure the clustering performance. Given a micro-array dataset containing N samples that belong to K classes, we assume that K is given in all the algorithms tested here. For each sample, x_n, let c_n be the cluster label predicted by an algorithm and r_n be the cancer type label provided by the dataset. The accuracy of the algorithm is defined as:

$$Accuracy = \frac{\sum_{n=1}^{N} I(r_n, c_n)}{N}, \quad (35)$$

where $I(A, B)$ returns 1 if $A = B$ and 0 otherwise.

Tested methods

We first compared the MCC with other loss functions between X and HW for the NMF algorithm on the cancer clustering problem, including l_2 norm distance, KL distance [10], α-divergence [4], and earth mover's distance (EMC) [12]. We further compared the proposed NMF-MCC algorithm with other NMF-based algorithms, including the penalized matrix decomposition

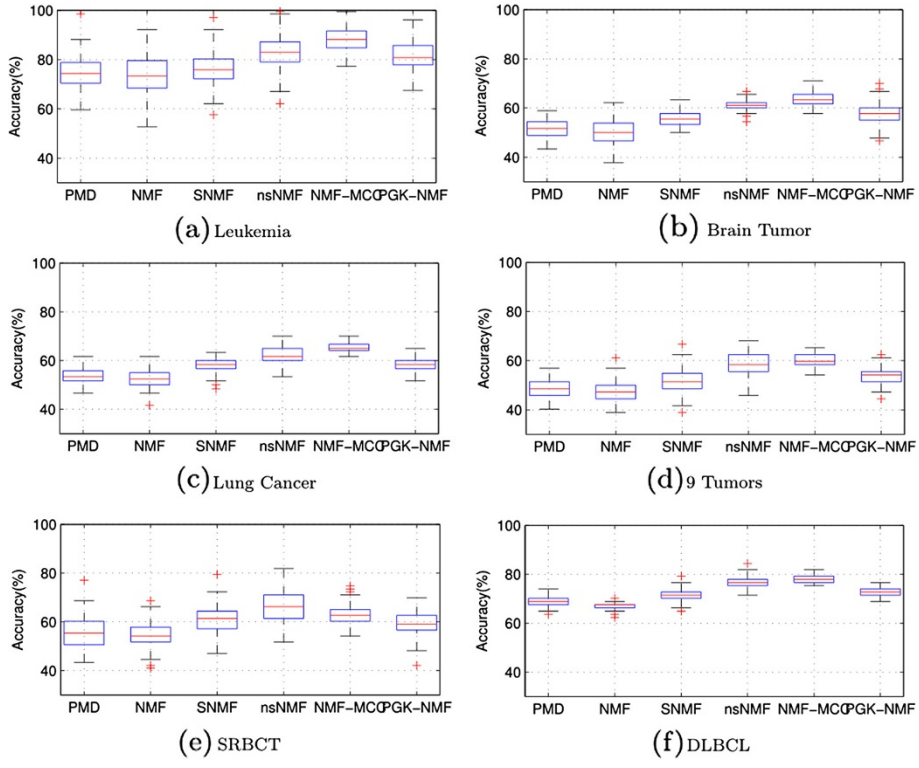

Figure 4 The boxplots of the clustering accuracies for different versions of NMF algorithms over 100 runs on the six gene expression datasets: (a) Leukemia, (b) Brain Tumor, (c) Lung Cancer, (d) 9 Tumors, (e) SRBCT, (f) DLBCL.

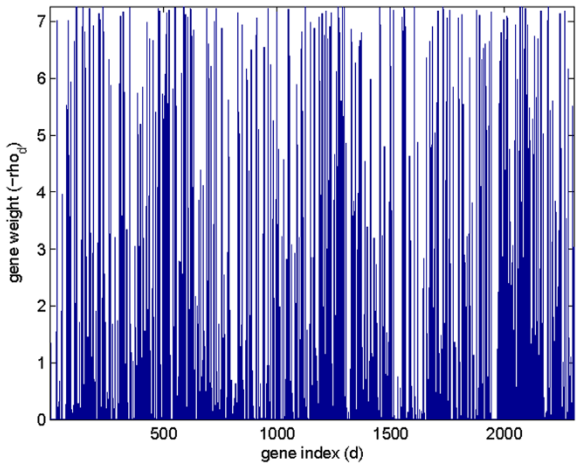

Figure 5 The gene weight vector learned by NMF-MCC with $-\rho$ on the SRBCT dataset.

(PMD) algorithm [7], the original NMF algorithm [22], the sparse non-negative matrix factorization (SNMF) algorithm [3], the non-smooth non-negative matrix factorization (nsNMF) algorithm [9] and the projected gradient kernel nonnegative matrix factorization (PGK-NMF).

Results

Since the initial H and W are selected randomly, we performed 100 independent trials and computed the average and the standard deviations of the accuracy for each loss function. The results from the comparison of MCC with other loss functions are presented in Figure 3. As shown in Figure 3, MCC consistently performed the best on all the six datasets. The other loss functions performed well on some datasets, but poorly on the others. It seems that the improvement of MCC increased when the number of genes increased. The standard deviation on the accuracy of MCC was much smaller than the standard deviation on the other loss functions, indicating that MCC is the most stable. On the other hand, EMD, although worked quite well in computer vision tasks [12], it did not perform well on gene expression data due to the significant difference between the image data and the gene expression data.

The results of the comparison of NMF-MCC with other related NMF methods are presented in Figure 4. Figure 4 shows the performance of different algorithms on the six datasets. The NMF-MCC algorithm outperformed the other algorithms on five out of the six datasets. The NMF-MCC algorithm could correctly cluster more than 88% and 78% of the samples in the Leukemia and DLBCL datasets, respectively, in a completely unsupervised manner. In contrast, the l_2 norm distance-based NMF algorithm performed even worse than the baseline PMD algorithm on the Leukemia and DLBCL datasets, i.e., an average accuracy of 73% and 67%, respectively. This verifies that correntropy is a much better measure of cancer clustering data. Note that NMF-MCC significantly outperformed the other algorithms on the Lung Cancer dataset, which contains a large number of genes. This implies that among the large number of genes, only a small fraction is likely to be relevant to cancerous tumor growth or spread. In NMF-MCC, the auxiliary variables $-\boldsymbol{\rho}$ acts as the feature selectors, we was able to select the relevant genes. Although the SNMF and nsNMF algorithms also improved on the performance of the baseline NMF algorithm, the improvement was much less than that of the NMF-MCC algorithm. A possible reason is that many genes exhibit similar patterns across all of the samples with only a few genes differentiating different cancer classes. They are likely to be sampled from a nonlinear manifold. Hence, the loss function defined by a linear kernel with either the

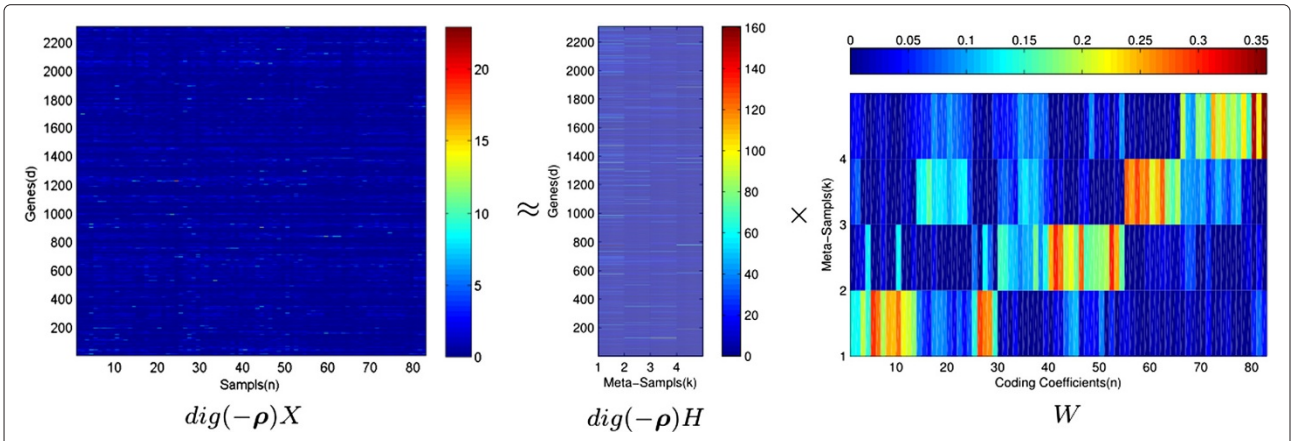

Figure 6 The meta-sample matrix, H, weighted by $dig(-\rho)$ and the corresponding coding matrix, W, obtained from the NMF-MCC algorithm for the SRBCT dataset.

l_2 norm or the KL distance could not capture them. In contrast, the NMF-MCC algorithm had a loss function that was defined by the correntropy and a Gaussian kernel, which could capture the nonlinear manifold structure much more effectively. By mapping the gene expression data into the nonlinear dataspace by a Gaussian kernel, the PGK-NMF outperformed the original NMF. However, our NMF-MCC could even further improve the PGK-NMF by applying different kernels to different features.

To understand what genes were selected by the NMF-MCC algorithm, we drew the gene weight figure on the SRBCT dataset (Figure 5). It can be seen that the $-\boldsymbol{\rho}$ vector is sparse, which shows the significance of certain genes. The resulting meta-sample matrix weighted by $-\boldsymbol{\rho}$ with the corresponding coding matrix is shown in Figure 6. By comparing to the coding matrix learned by the original NMF with the l_2 norm distance in Figure 1, we determine that the coding matrix learned by the NMF-MCC algorithm is much more discriminative among different cancer classes. On this dataset, the NMR-MCC algorithm achieved an average clustering accuracy of 63%.

Discussion

Traditional unsupervised learning techniques select features with features selection algorithms and then do clustering using the selected features. The NMF-MCC algorithm proposed here achieves both goals simultaneously. The learned gene weight vector reflects the importance of the genes in the gene clustering task, and the coding matrix encodes the clustering results for the samples.

Our experimental results demonstrate that the improvement of NMR-MCC over the other methods increases when the number of genes increases. This shows the ability of the proposed algorithm to effectively select the important genes and cluster samples. This is an important property because high-dimensional data analysis has become increasingly frequent and important in diverse fields of sciences and engineering, and social sciences, ranging from genomics and health sciences to economics, finance and machine learning. For instance, in genome-wide association studies, hundreds of thousands of SNPs are potential covariates for phenotypes such as cholesterol level or height. The large number of features presents an intrinsic challenge to many classical problems, where usual low-dimensional methods no longer apply. The NMF-MCC algorithm has been demonstrated to work well on the datasets with small numbers of samples but large numbers of features. It can therefor provide a powerful tool to study high-dimensional problems, such as genome-wide association studies.

Conclusion

We have proposed a novel NMF-MCC algorithm for gene expression data-based cancer clustering. Experiments demonstrate that correntropy is a better measure than the traditional l_2 norm and KL distances for this task, and the proposed algorithm significantly outperforms the existing methods.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JW designed and implemented the algorithm, conducted the experiments, performed data analysis and drafted the manuscript. XW revised the manuscript. XG supervised the study and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The study was supported by a grant from King Abdullah University of Science and Technology, Saudi Arabia. We would like to thank Dr. Ran He for the discussion about the maximum correntropy criterion at ICPR 2012 conference.

Author details

Received: 16 February 2012 Accepted: 8 March 2013
Published: 24 March 2013

References

1. Shi F, Leckie C, MacIntyre G, Haviv I, Boussioutas A, Kowalczyk A: **A bi-ordering approach to linking gene expression with clinical annotations in gastric cancer.** *BMC Bioinformatics* 2010, **11:**477.
2. de Souto MCP, Costa IG, de Araujo DSA, Ludermir TB, Schliep A: **Clustering cancer gene expression data: a comparative study.** *BMC Bioinformatics* 2008, **9:**497.
3. Gao Y, Church G: **Improving molecular cancer class discovery through sparse non-negative matrix factorization.** *Bioinformatics* 2005, **21**(21):3970—3975.
4. Liu W, Yuan K, Ye D: **On alpha-divergence based nonnegative matrix factorization for clustering cancer gene expression data.** *Artif Intell Med* 2008, **44**(1):1–5.
5. Zheng CH, Ng TY, Zhang L, Shiu CK, Wang HQ: **Tumor classification based on non-negative matrix factorization using gene expression data.** *IEEE Trans Nanobioscience* 2011, **10**(2):86–93.
6. Kim MH, Seo HJ, Joung JG, Kim JH: **Comprehensive evaluation of matrix factorization methods for the analysis of DNA microarray gene expression data.** *BMC Bioinformatics* 2011, **12**(Suppl 13):S8.
7. Zheng CH, Zhang L, Ng VTY, Shiu SCK, Huang DS: **Molecular pattern discovery based on penalized matrix decomposition.** *IEEE/ACM Trans Comput Biol Bioinformcs* 2011, **8**(6):1592–1603.
8. Tjioe E, Berry M, Homayouni R, Heinrich K: **Using a literature-based NMF model for discovering gene functional relationships.** *BMC Bioinformatics* 2008, **9**(7):P1.
9. Carmona-Saez P, Pascual-Marqui R, Tirado F, Carazo J, Pascual-Montano A: **Biclustering of gene expression data by non-smooth non-negative matrix factorization.** *BMC Bioinformatics* 2006, **7:**78.
10. Venkatesan R, Plastino A: **Deformed statistics Kullback-Leibler divergence minimization within a scaled Bregman framework.** *Phys Lett A* 2011, **375**(48):4237–4243.
11. Cai D, He X, Han J, Huang TS: **Graph regularized nonnegative matrix factorization for data representation.** *IEEE Trans Pattern Anal Mach Intell* 2011, **33**(8):1548–1560.
12. Sandler R, Lindenbaum M: **Nonnegative matrix factorization with earth mover's distance metric for image analysis.** *IEEE Trans Pattern Anal Mach Intell* 2011, **33**(8):1590–1602.
13. He R, Zheng WS, Hu BG: **Maximum correntropy criterion for robust face recognition.** *IEEE Trans Pattern Anal Mach Intell* 2011, **33**(8):1561–1576.
14. Zafeiriou S, Petrou M: **Nonlinear nonnegative component analysis.** In *CVPR: 2009 IEEE Conference on Computer Vision and Pattern Recognition,*

Vols 1-4. Miami: IEEE Conference on Computer Vision and Pattern Recognition; 2010:2852–2857.
15. Yan H, Yuan X, Yan S, Yang J: **Correntropy based feature selection using binary projection.** *Pattern Recognit* 2011, **44**(12):2834–2842.
16. He R, Hu BG, Zheng WS, Kong XW: **Robust principal component analysis based on maximum correntropy criterion.** *IEEE Trans Image Process* 2011, **20**(6):1485–1494.
17. Chalasani R, Principe JC: **Self organizing maps with the correntropy induced metric.** In *Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN2010)*. Barcelona, Spain; 2010:1–6.
18. Liu W, Pokharel PP, Principe JC: **Correntropy: properties and applications in non-gaussian signal processing.** *IEEE Trans Signal Process* 2007, **55**(11):5286–5298.
19. Horaud R, Forbes F, Yguel M, Dewaele G, Zhang J: **Rigid and articulated point registration with expectation conditional maximization.** *IEEE Trans Pattern Anal Mach Intell* 2011, **33**(3):587–602.
20. BEER G: **Conjugate convex-functions and the epi-distance topology.** *Proc Am Math Soc* 1990, **108**(1):117–126.
21. Qi Y, Ye P, Bader J: **Genetic interaction motif finding by expectation maximization - a novel statistical model for inferring gene modules from synthetic lethality.** *BMC Bioinformatics* 2005, **6:**288.
22. Lee DD, Seung HS: **Algorithms for non-negative matrix factorization.** *Adv Neural Inf Process Syst* 2001, **13:**556–562.
23. Statnikov A, Aliferis C, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.** *Bioinformatics* 2005, **21**(5):631–643.
24. Shipp M, Ross K, Tamayo P, Weng A, Kutok J, Aguiar R, Gaasenbeek M, Angelo M, Reich M, Pinkus G, Ray T, Koval M, Last K, Norton A, Lister T, Mesirov J, Neuberg D, Lander E, Aster J, Golub T: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8**(1):68–74.
25. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E: **Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**(5439):531–537.
26. Pomeroy S, Tamayo P, Gaasenbeek M, Sturla L, Angelo M, McLaughlin M, Kim J, Goumnerova L, Black P, Lau C, Allen J, Zagzag D, Olson J, Curran T, Wetmore C, Biegel J, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis D, Mesirov J, Lander E, Golub T: **Prediction of central nervous system embryonal tumour outcome based on gene expression.** *Nature* 2002, **415**(6870):436–442.
27. Bhattacharjee A, Richards W, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark E, Lander E, Wong W, Johnson B, Golub T, Sugarbaker D, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci* 2001, **98**(24):13790–13795.
28. Staunton J, Slonim D, Coller H, Tamayo P, Angelo M, Park J, Scherf U, Lee J, Reinhold W, Weinstein J, Mesirov J, Lander E, Golub T: **Chemosensitivity prediction by transcriptional profiling.** *Proc Natl Acad Sci* 2001, **98**(19):10787–10792.
29. Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C, Meltzer P: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7**(6):673–679.

doi:10.1186/1471-2105-14-107
Cite this article as: Wang *et al.*: Non-negative matrix factorization by maximizing correntropy for cancer clustering. *BMC Bioinformatics* 2013 **14**:107.