Semi-nonnegative Matrix Factorization Algorithm Based on Genetic Algorithm Initialization

M. Chouh and K. Boukhetala

***Abstract*—Semi-nonnegative matrix factorization (Semi-NMF) is one of variations of nonnegative matrix factorization model (NMF) when the data matrix *X* is unconstrained (it may have mixed signs). Semi-NMF decomposes *X* into two matrices A and *B* of dimensions $n \times k$ and $k \times p$ respectively, where each element of the matrix *B* is nonnegative, such that: $X \approx AB$. In the present paper, we proposed a semi-nonnegative matrix factorization algorithm based on genetic algorithm (GA) initialization which has larger searching area and gives the best initialization for the Semi-NMF algorithm to get the optimal solution of semi-nonnegative matrix factorization problem. Also, we compared this initialization for Semi-NMF algorithm with both the random and the k-means initializations introduced in the literature.**

***Index Terms*—Semi-nonnegative matrix factorization, genetic algorithm, initialization.**

I. Introduction

Many applications in computer science and intelligent systems include analysis of large scale and high dimensional data. When collections of extensive information are dealt, it is usually very computationally expensive to perform some operations on the row (resp. on the column) form of the data. Therefore, suitable methods approximating the data in lower dimensions or with lower rank are needed. Matrix factorization (or matrix decomposition) is an important task in data analysis and processing which attracted the dimensionality reduction of the data.

Several approaches of matrix factorization have been developed over many decades. For instance, using dominant subspaces with the singular value decomposition (SVD) has been proposed as the best model to reduce the complexity of data and complicated systems which underlies principal component analysis (PCA) [1]: a statistical technique that has found application in fields such as face recognition and image compression, and it is a common technique for finding patterns in data of high dimension.

In the context of nonnegative data, models such as PCA cannot give ideal results for interpretation, for this reason, one low-rank approximation technique with additional nonnegativity constraints was proposed by Lee and Seung in 1999 and it is referred to as nonnegative matrix factorization (NMF) [2].

In particular, nonnegative matrix factorization (NMF) is recently very popular unsupervised learning algorithm for efficient factorization of real data matrices implementing the nonnegativity constraint [3]. In this work, we focus on one variant of NMF algorithms which is the Semi-nonnegative matrix factorization (Semi-NMF) proposed by Ding, C. in 2006 [4] and used successfully in several applications, by [5] for motion segmentation with missing data, by [6] for image super resolution, and by [7] for hyperspectral unmixing.

Semi nonnegative matrix factorization can be defined as follows: Given a matrix $X \in \Re^{n \times p}$ and a factorization rank k, solve:

$$\min_{A \in \Re^{n \times k}, B \in \Re^{k \times p}} \left\| X - AB \right\|_F \text{ such that } B \geq 0 \quad (1),$$

where $\| \ \|_F$ is the Frobenius norm, and $B \geq 0$ means that all components of B are nonnegative.

C. Ding was introduced the semi-nonnegative matrix factorization problem as a relaxation of the clustering context named the k-means by allowing the matrix B defined in (1) (binary matrix named as the cluster indicator matrix in the k-means task) to be a nonnegative matrix (not necessarily a binary matrix).

In the paper [8], we were attracted the problem of exact semi-nonnegative matrix factorization, where the factorization form of the data matrix X is the following:

$$X = AB \ , A \in \Re^{n \times k}, \ B \in \Re_{+}^{k \times p} \quad (2).$$

The smallest integer k for which the factorization (2) can be obtained is defined as the semi-nonnegative rank for the matrix X. the study of this rank sheds additional light to the introduced approximation problem and vice versa.

The main deference between the factorization (2) and Ding's approximation problem is that the first does not focus on optimal approximation of X by AB, but it requires that the factorization gives a perfect fit of X. Also, the integer k is given (known) in Ding's approximation problem but it is unknown in the semi-nonnegative rank problem (2). Although, these two problems are different.

GA is a classical method in data mining and machine

Manuscript received May 10, 2016; revised August 2, 2016.

The authors are with Faculté de mathématiques, USTHB, El-Alia BP 32, Bab-Ezzouar 16111, Alger, Algérie (merich_88@hotmail.com, kboukhetala@usthb.dz).

doi: 10.18178/ijmlc.2016.6.4.603

learning, it is an optimization and search technical based on the principles of biological evolutions, genetics and natural selections. It is acknowledged as good solver for tough problems [9]. Some of advantages of the GA include that it:

1) Optimizes with any type of variables (discrete or continuous).
2) Optimizes without taking account of the complexity of the objective function.
3) Is well suited for parallel implementation.
4) And the more interest advantage is that the GA has a large searching area and can reach the global optimal solution of the optimization problem.

Genetic algorithms has widely used in the field of matrix factorization. For example: an initial version of genetic algorithms was adapted for binary matrix factorization in [10], in the paper [11], the application of five population based algorithms (where the genetic algorithm appear) as new initialization variant for nonnegative matrix factorization is presented, and in [12], we can find an improved nonnegative matrix factorization algorithm based on genetic algorithm.

The paper is structured as follows: In the next section, the Semi-NMF algorithm is presented briefly. Section III explains how to work the genetic algorithm. In section IV, we present our proposed genetic algorithm for Semi-NMF initialization. Section V explains how the parallel computing can save in the algorithmic complexity introduced by the iterations the initialization algorithm. In section VI, we show with an experimental study how the performance of our proposed algorithm is get. Finally, we present some conclusions in section VII.

II. The Semi-NMF Method

Ding was proposed an efficient algorithm for resolving the Semi-NMF problem (1) which is an iterative updating algorithm that alternatively updates the matrices A and B by the following Alg.1:

begin
initialize B by doing the k-means clustering. %It gives cluster indicators matrix B: $B_{ij}=1$ if and only if the j^{th} column of X belongs to the ith cluster and $B_{ij}=0$ otherwise.
Update A using the rule:

$$A = XB(BB^t)^{-1}$$

%note that B is a positive semidefinite matrix, then the inversion of this matrix is trivial.
Update B using the rule:

$$B_{ij} \leftarrow B_{ij}\sqrt{\frac{\max(0,(A\ X)_{ij})+\max(0,(A^tAX)_{ij})}{\max(0,-(A\ X)_{ij})+\max(0,(A^tAX)_{ij})}}$$

until
termination-condition is satisfied
end

Alg. 1. Semi-NMF multiplicative update algorithm.

Ding utilized the gradient descent to update the matrix the matrix B by choosing a smart step size. For the implementation purpose (using Matlab), a small constant is added to all elements of B to avoid division by zero.

III. Genetic Algorithm

A GA searches through a space of «chromosomes », each of which represents a candidate solution to a given problem (in some cases, a solution consists of a set of chromosomes). Most methods called GAs have at least the following elements in common: populations of chromosomes, selection according to the fitness function (objective function to optimize), crossover to produce new offspring, and random mutation of new offspring.

The general scheme of a GA can be given as follows:

begin
initialize population with random candidate solutions;
evaluate each candidate;
repeat
select parents;
recombine pairs of parents;
mutate the resulting children;
evaluate children;
select individuals for the next generation
until
termination-condition is satisfied
end

Alg. 2. Genetic algorithm structure.

There are variants of standard generational GA. The differences are mostly in particular selection, crossover, mutation and replacement strategy [13].

IV. The Proposed Initialization Algorithm For Semi-Nonnegative Matrix Factorization

The goal of our initialization algorithm is to find heuristically optimal starting points for single columns of the matrix B, which can be computed with the following algorithm:

Given a matrix $X \in \Re^{n\times p}$ and $k << \min(n, p)$
$A0 = randn(n,k)$ % a random initialization for the matrix A
for i=1 to p
Use the GA to minimize $\left\|X_j - A_0 B_j\right\|_F$;
$B(:, j) = B_j$.
end

Alg. 3. Semi-NMF initialization algorithm.

The input parameters for the genetic algorithm are X_j (the j^{th} column of the original matrix X) and A_0. the output is the initialized column vector B_j (the j^{th} column of B).

To keep the nonnegativity of elements of the matrix B, GA takes upper/lower bounds to the search space.

V. Parallel Computing

The call of the genetic algorithm for each column of the matrix B can make some algorithmic complexity when the dimensionality of B is too large, and since the initialization of any column of B does not influence the initialization of any other column of B. This allows for a parallel implementation of the proposed initialization method, i.e., the

columns of B can be computed in parallel.

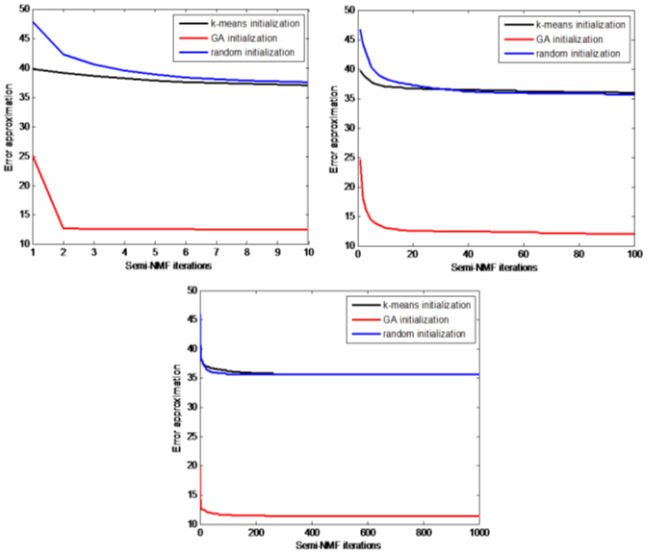

Fig. 1. Error approximation of the Semi-NMF multiplicative update algorithm with different initializations on the Ionosphere data set (case: k=5).

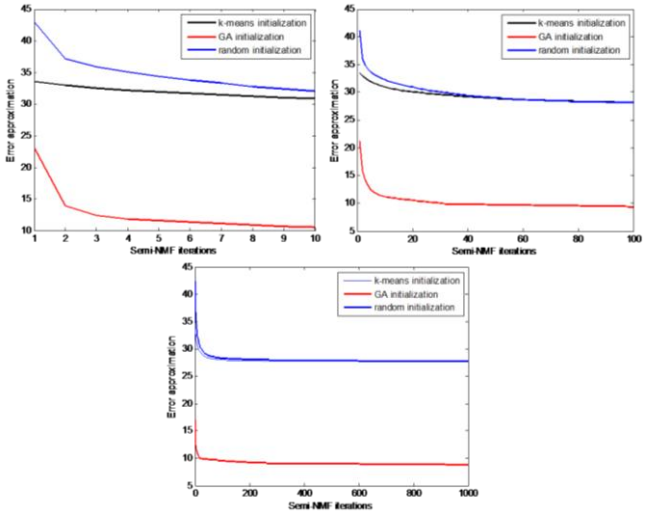

Fig. 2. Error approximation of the Semi-NMF multiplicative update algorithm with different initializations on the Ionosphere data set (case: k=10).

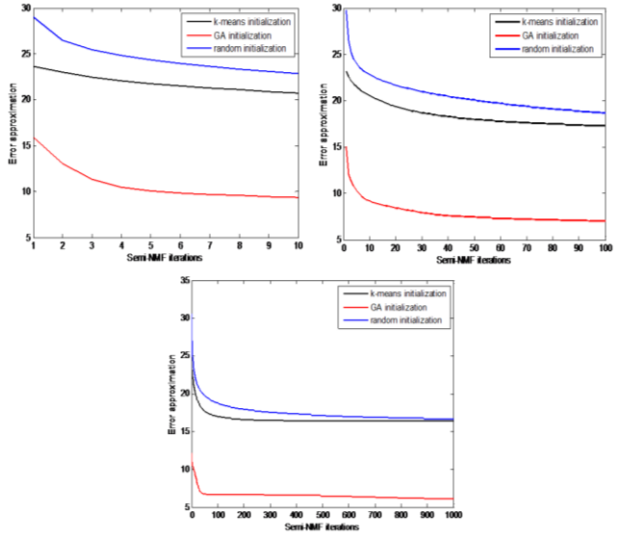

Fig. 3. Error approximation of the Semi-NMF multiplicative update algorithm with different initializations on the Ionosphere data set (case: k=20).

VI. Numerical Experiments

We present in this section the results of a simple experiment on Ionosphere UCI data set (it is a 34 by 351 matrix that contain both positive and negative entries in the interval [-1, 1]) which aims to verify that the use of genetic algorithm in the initialization of the Semi-NMF algorithm sheds additional light to the Semi-NMF algorithm. We was adapted the GA implementation existed in Matlab (ga) for doing the initialization algorithm.

Our goal is to illustrate on real data, the observations about the error approximation for three Semi-NMF initializations (random initialization, k-means initialization, and GA initialization).

In figures: Fig. 1, Fig. 2 and Fig. 3, we show that a genetic algorithm initialization for the multiplicative update variant of the Semi-NMF algorithm give a lower approximation error function than the random and the k-means initializations.

VII. Conclusion

It is known that the genetic algorithm with the real coding has not a theoretical improvement, and we cannot verify the optimization conditions for a GA, but in the world of computer sciences, it gives the best results, and it is favorite by the users.

The use of GA is independent with the Semi-NMF algorithm implementation or it has not modified his quality of convergence because the GA is used just in the algorithm's initialization. In the other hand, the GA initialization gives a best starting points for the Semi-NMF multiplicative update algorithm.

References

[1] G. Golub and C. Van Loan, *Matrix Computation*, 3rd Edition, The Johns Hopkins University Press, 1996, pp. 69-74.
[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.
[3] D. Chen and R. J. Plemmons, "Nonnegativity constraints in numerical analysis," in *Proc. Symposium on the Birth of Numerical Analysis*, 2007, pp. 109-140.
[4] C. Ding, T. Li, and M. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45–55, 2010.
[5] Q. Mo and B. Draper, "Semi-nonnegative matrix factorization for motion segmentation with missing data," *Computer Vision-ECCV*, pp. 402-415, 2012.
[6] M. Bevilacqua, A. Roumy, C. Guillemot *et al*., "Neighbor embedding based single-image super-resolution using semi-nonnegative matrix factorization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 1289-1292.
[7] N. Y. Koya, J. Chanussot, and A. Iwasaki, "Generalized bilinear model based nonlinear unmixing using Semi-nonnegative matrix factorization," in *Proc. IEEE Int. Geoscience and Remote Sensing Symposium*, 2012, pp. 1365-1368.
[8] M. Chouh, M. Hanafi, and K. Boukhetala, "Semi-nonnegative rank for real matrices and its connection to the usual rank," *Linear Algebra and Its Applications*, vol. 466, pp. 27–37, 2015.
[9] J. Gareth, "Genetic and evolutionary algorithms," E*ncyclopedia of Computational Chemistry*, John Wiley and Sons, 1998.
[10] V. Snasael, J. Platos, and P. Krömer, "Developing genetic algorithms for boolean matrix factorization," *DATESO'2008*, 2008, pp. 61-70.
[11] A. Janecek and Y. Tan, "Using population based algorithms for initializing nonnegative matrix factorization," *Advances in Swarm Intelligence*, pp. 307-316, 2011.
[12] Z. Sheng, Y. Zhi, and W. Can, "An improved non-negative matrix factorization algorithm based on genetic algorithm," in *Proc. ICCSET*,

Int. Conf. on Computer Science and Electronic Technology, 2014, pp. 395-398.
[13] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*, 2nd ed. John Wiley and Sons, 2005.

Kamel Boukhetala is the Dean of the Faculty of Mathematics-USTHB, Algiers, Algerian Adviser to Finance Ministry, Adviser to Ministry of Statistics & Forecasting. Professor Boukhetala received his Ph.D. in mathematics, option: operational research statistics and probability from Joseph Fourier University – Grenoble. His research interests are stochastic processes, mathematical and computational finance, stochastic optimization, bayesian statistics and decision theory, stochastic models in finance and actuarial. He has published widely in leading academic journals, research books and author of the Sim.Diffproc and Sim.Diffprogui R-packages. He is an invited reviewer in many journals.

Meriem Chouh is a PhD student at the Department of probability and statistics, University of Houari Boumedienne of Algeria. She is preparing her doctoral thesis in the field of factorization performance algorithms, her work is centered around developing and integrating innovative statistical approaches based on the application of factorization models in the theory of data analysis. Her research interests include: probability and statistics, stochastic modeling, simulation theory, the insurance, financial and actuarial mathematics.