An improved non-negative matrix factorization algorithm based on genetic algorithm

Sheng Zhou[1,a], Zhi Yu[1,b] and Can Wang[1,c]

[1]Zhejiang Provincial Key Laboratory of Service Robot,College of Computer Science, Zhejiang University, Hangzhou 310027, China

[a]zhousheng_zju@zju.edu.cn, [b]yuzhirenzhe@zju.edu.cn, [c]wcan@zju.edu.cn

Keywords: Non-negative matrix factorization, Genetic algorithm, Document clustering, Robustness

Abstract. The non-negative matrix factorization (NMF) algorithm is a classical matrix factorization and dimension reduction method in machine learning and data mining. However, in real problems, we always have to run the algorithm for several times and use the best matrix factorization result as the final output because of the random initialization of the matrix factorization. In this paper, we proposed an improved non-negative matrix factorization algorithm based on genetic algorithm (GA), which uses the internal parallelism and the random search of genetic algorithm to get the optimal solution of matrix factorization. It could have larger searching area and higher accuracy in matrix factorization. In the document clustering problem, we use the TDT2 dataset and design several contrast experiments on the classical NMF and the improved NMF based on genetic algorithm, the experiment results show that our improved non-negative matrix factorization algorithm has higher clustering accuracy and better robustness.

Introduction

With the rapid development of the Internet and the storage technology, we have to process very large scale of data, which is often in the form of sparse matrix. While processing the data, matrix factorization has been a particularly useful tool to reduce matrix decompositions. Among different matrix factorization algorithms, the non-negative matrix factorization (NMF) algorithm has been widely usedfor its special nonnegative constraint, factorization results can be interpreted and the simplicity of the factorization process. With the deepening of the research, it has been successfully applied to various fields. [1]

However, considering the randomness of the initial matrix in the real problem, we usually have to use NMF to do factorization for several times on the same objective matrix and choose the best factorization result among them as the final result. This will absolutely bring the waste of computation time since the matrix in different iteration process is mutual independent. This gives us enough space to improve the performance of the algorithm in the process of applying NMF algorithm in the real problem.

The classical genetic algorithm use ideas based on the language of natural genetics and biological evolution. [2] It is a classical algorithm in data mining and machine learning. The best advantage of the genetic algorithm is that it has a large searching area and can reach the global optimal solution of the problem. Also, it does not need complex computation and has rapid convergence, as a result of which, the genetic algorithm is suitable for handling the complex and non-linear problem that the traditional algorithm can’t solve. In this paper, we proposed a new non-negative matrix factorization algorithm based on genetic algorithm (GA) to strengthen the correlation and expand the search domain.

Rather than using NMF for several times, we apply the population evolution processing to the iteration of the NMF: 1.We consider the row of the decomposition matrix as the gene of the individual. 2. We consider a complete iteration of NMF as a process of the evolution of the individual. According to simulating the process of population evolution, we can make full use of every matrix during the iteration and extend the searching space so that we can accept better factorization result.

Non-negative matrix factorization

Non-negative matrix factorization is a classical matrix factorization method,which was first proposed by Lee and Seung. Different from the other classical matrix factorization method such as Principal Component Analysis(PCA), Singular Value Decomposition (SVD), the non-negativity constraints make the represention purely additive(aloowing no subtractions) and the non-negative matrix factorization algorithm has been widely used in different fields: document clustering[3], image analysis[4][5], voice recognition. However, in the process of using the non-negative matrix factorization, we have to face the fact that we have to run the algorithm for several times to reduce the side effect of the randomly initialization. During the use of the non-negative matrix factorization, we find that the matrix of different run-times are totally independent, this means that we can improve the performance of matrix factorization using the hidden information of the matrix among the iterations of NMF.

The proposed method

In this section, we will first come up with the basic assumptions and propose the non-negative matrix factorization based on the genetic algorithm (GANMF).

In order to combine the NMF problem with the GA problem, we have to process the elements in the NMF problem and make some assumptions. First, we assume that every row of the decomposition matrix represents the gene information of the individual and the decomposition matrix itself is an individual. Next, we assume that a complete iteration of NMF is the process of the evolution of the individual. Above all, we can preliminary combine the NMF problem with the GA problem.

After the above operations, we will propose the specific process of the non-negative matrix factorization based on genetic algorithm (GANMF). Given a non-negative matrix V, the purpose of NMF is to find non-negative matrix factors W and H such that: [3]

$$V \approx WH \tag{1}$$

The goal is to minimize the following object function:

$$J = \frac{1}{2}||V - WH||^2 \tag{2}$$

Using the Lagrange multiplier method, we can get the update rules as follows:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T V)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} \tag{3}$$

$$W_{\alpha\mu} \leftarrow W_{\alpha\mu} \frac{(VH^T)_{\alpha\mu}}{(WHH^T)_{\alpha\mu}} \tag{4}$$

In our GANMF method, we will use the classical NMF algorithm for **t**times. Each time of the using, we name it as a individual, in view of the biotic population in the nature. That means each individual is a completely process of the classical non-negatibe matrix factorization, which is a combination of a several iterations.

When we get the m × n non-negative matrix V, which needs to be factorizated, we first randomly create the initial m × kmatrix W andn × kmatrix H for every population, that will be **t** pairs of random matrix. For these random matrix, we use the update rules (3)(4) one time to update the matrix W and H, which gives us **t**pairs of new matrix W and H. The classical non-negative matrix factorization does the updating operations until we reach the condition of convergence. But in our GANMF algorithm, every time we update the matrix W and H, we apply the genetic operator: selection, crossover and mutation on the matrix W and H, which will gives us a bigger searching area. However, simply expansion the searching area will bring large computation complexity and the improvement of the matrix factorization may not be worthy compared with the increase of the computation complexity, thus we use the optimal choice policy to keep the scale of the population and reduce the computation complexity at the same time.After the optimal choice operation, we get the best individuals for the object funtion(2). Then we apply the update rules on the individuals again and stop until reaching the condition of convergence. The flow chart is as follows:

In summary, our algorithm is composed of the following steps:

Experiment

To conduct the performance of our improved non-negative matrix factorization algorithm in document clustering, weuse the TDT2 document corpora, which has been generally acceptedin document clustering.To form the testing data, we select documents from different clusters randomly picked from the TDT2.In order to measure the effect of our algorithm in document clustering problem, we use two indexes which is the accuracy(AC) and the normalized mutual information ($\widehat{MI}$). The AC is defined as follows:

$$AC = \frac{\sum_{i=1}^{n} \delta(\alpha_i, map(l_i))}{n} \tag{5}$$

where n denotes the total number of documents in the test,$\delta(\mathrm{x}, \mathrm{y})$ is the delta function that equals one if x = y andequals zero otherwise, and $\mathrm{map}(l_i)$ is the mapping functionthat maps each cluster label li to the equivalent label fromthe document corpus.Given two clusters A and B, the MI is defined as follows:

$$MI(A,B) = \sum_{a_i \in A, b_i \in B} p(a_i, b_i) \cdot log_2 \frac{p(a_i, b_i)}{p(a_i) \cdot p(b_i)} \tag{6}$$

Where $\mathrm{p}(a_i)$ $p(b_i)$ denote the probabilities that a document belongs to the cluster A and B, and $\mathrm{p}(a_i, b_i)$denotes the jiont probability that the document belongs to the cluster A and B at the same time. To normalize the MI which takes values between zero and one, we use $\widehat{MI}$ as follows:

$$\widehat{MI}(\mathrm{A}, \mathrm{B}) = \frac{MI(A,B)}{\max(H(A), H(B))} \tag{7}$$

To reduce the side effect of randomly selection, we repeat the experiment for several times and compare the result of both algorithms on the document clustering according to the mean and variance.

Table 1 Performanc comparisons of NMF and GANMF

t	AC		$\widehat{MI}$	
	GANMF	NMF	GANMF	NMF
1	0.6384	0.6328	0.0244	0.0168
2	0.7926	0.8140	0.1906	0.2104
3	1	1	1	1
4	0.9601	0.8659	0.7963	0.5362
5	0.9874	0.9874	0.9097	0.9097
6	0.6904	0.6904	0.0049	0.0049
7	0.6893	0.6893	0.0102	0.0186
8	0.9954	0.9954	0.9576	0.9576
9	1	1	1	1
10	0.7841	0.7841	0.1652	0.0778
average	0.8538	0.8246	0.5059	0.4732
variance	0.1493	0.1688	0.4575	0.4526

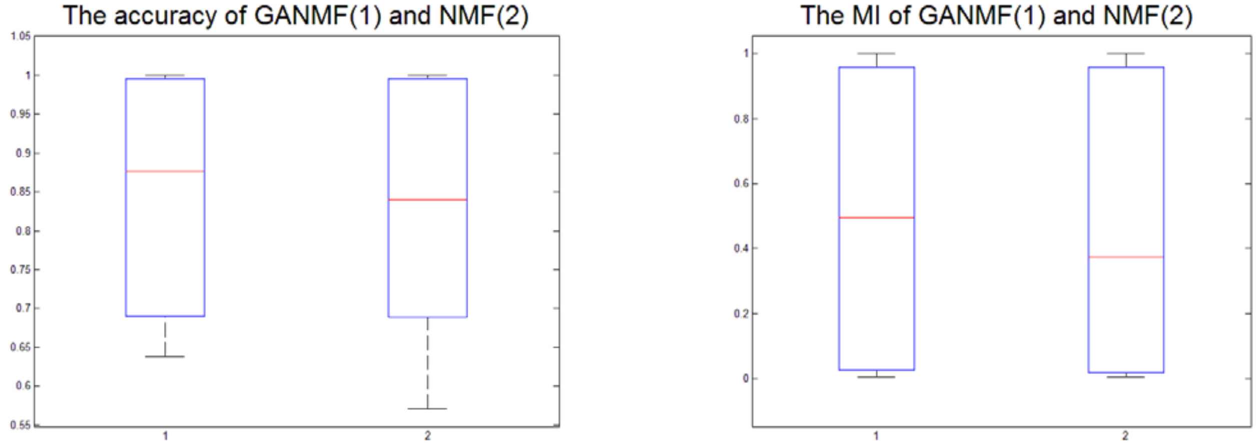

Figure 1 The comparison of the AC and $\widehat{MI}$ between GANMF and NMF

Table 1 and Figure 1 show the result of two algorithms, GANMF and NMF, applying on ten different randomly selected clusters from TDT2 document corpora. For each test cluster, we apply the algorithm for 50 runs and the final result is obtained by averaging the results from 50 runs. From the table and the figure, it is obvious to know that for the accuracy of the clustering and the $\widehat{MI}$ between the clusters, GANMF algorithm has better results than the NMF algorithm, and the robustness of GANMF is a little better than the NMF algorithm.

Conclusion

In this paper, we have proposed an improved non-negative matrix factorization algorithm based on genetic algorithm. Different from the classical non-negative matrix factorization algorithm, the GANMF make full use of the hidden information among the matrixes during the process of the iteration, as a result of which, GANMF has a larger searching area and a better robustness than the NMF algorithm. As evidenced by the experiment, our GANMF algorithm has a better performance in document clustering.

Acknowledgement

This work is supported byNational Key Technology R&D Program(Grant No. 2012BAI34B01)

References

[1] Hyunsoo.kim. Haesun.Park,:Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method.

[2] Garrett.M.Morris, David.S.Goodsell, Ruth.Huey: Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function.

[3] Wei Xu, Xin Liu, Yihong Gong.: Document Clustering Based On Non-negative Matrix Factorization

[4] I.Buciu, I.Pitas, Sao Luis. In: A new sparse image representation algorithm applied to facial expression recognition. Proc of IEEE Workshop on Machine Learning for Signal Processing

[5] Tao Feng, Stan Z.Li, Heung-Yeung.: Local Non-negative Matrix Factorization as a Visual Representation.

[6] Bhuksha Raj, Tuomas Virtanen, SourishChaudhuri.: Non-negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition.