NEW BICLUSTERING MODEL USING BINARY MATRIX FACTORIZATION (BMF) FOR CLUSTERING OF THE ENVIRONMENTAL PROJECT RISKS

Najafi Asadallah
Department of Industrial Engineering (I.E.), Islamic Azad University, Zanjan Branch, Zanjan (IRAN)

Received December 05, 2012 Accepted March 07, 2013

ABSTRACT

It is commonly recognized that knowledge is the only source of core competence of Environmental Projects in the knowledge based companies, but the productivity rates of projects is always low. Clustering, as a widely used data mining approach, has been used to discover phenotypes from the raw expression data. However traditional clustering algorithms have limitations since they cannot identify the substructures of samples and features hidden behind the data. Different from clustering, biclustering is a new methodology for discovering genes that are highly related to a subset of samples. Several biclustering models/methods have been presented and used for environmental research. Based on projects characteristics, in this paper, we seek to identify factors influencing on the Environmental Projects Risks (EPR. finally, we seek to use the Binary Matrix Factorization (BMF) method for clustering of the Environmental Projects Risks (EPR). This paper will help Environmental Projects managers to implement different corresponding measures. A case study is presented where this model measures and validates at the lead and zinc Zanjan Company.

Key Words : Environmental Projects Risk (EPR), Knowledge, Clustering, Binary Matrix Factorization (BMF)

INTRODUCTION

Activities done in the fields of Environmental Projects Risk Management (EPRM) are in the early stages of its growth. One of the ways that EPRM can be successful is that it brings first make sure the Environmental Projects technically have the ability to run. Human resource management is therefore important in any organization. Environmental Projects are no different in this regard. Studies have done by the PMI Institute in 2007 identified more than 90 percent lead Environmental Projects implemented have adhered to this requirement, secondly factors affecting successful management Environmental Projects should be identified. Basic research has proven that human resources are the most important factor for Environmental Projects success. Nowadays, human resources position in Environmental Projects not only revised their strategic role in the successful management Environmental Projects has been gradually accepted, but have concluded that human factors over technical issues led to successful Environmental Projects. Despite these findings, only a small ratio of the number of empirical research have been done so far. In the past, Environmental Projects success has been analyzed which is based on three main factors, including cost, time and performance. One of the fundamental problems of the past approach is lack of attention to other aspects of the Environmental Projects. The study tried to examine the most important dimensions of Environmental Projects success such as human resources. Human resource management process contains the necessary coordination of human resources in the Environmental Projects. These processes include needs for designing, providing, advice, allocation of staff time and clearance Environmental Projects. All this process depends on the activities that assigned to them. Activities can be an intellectual and manual. Intellectual work
required due to specific knowledge will have a lot of complexity. Therefore, human resources involved in intellectual will spend more resources. If management is not suitable for the Environmental Projects affects output quality. The other hand, the increasing interest around knowledge worker and has caused a significant body of empirical research to emerge, examining the impact of different knowledge workers factors on EPRM success. However, minimum attention has been given to the conception or understanding of the specific strategies through which knowledgeable workers factors influence EPRM. Improving the productivity of knowledgeable workers is one of the most important challenges for companies that faces the transition from the industrial economy to an economy based on information and knowledge. Knowledgeable worker’s are obviously non-manual workers and are usually employed by Environmental Projects managers to carry out innovative activities. Knowledge worker is a member of the Environmental Projects organization who uses knowledge to be a more productive worker. A knowledgeable worker is anyone who works for a living at the tasks of developing or using knowledge. In this paper, we present a new biclustering model using Binary Matrix Factorization (BMF). BMF is a new variant rooted from non-negative Matrix Factorization (NMF). We begin by proving a new boundedness property of NMF. Two different algorithms to implement the model and their comparison are then presented. We show that the microarray data biclustering problem can be formulated as a BMF problem and can be solved effectively using our proposed algorithms. Unlike the greedy strategy-based algorithms, our proposed algorithms for BMF are more likely to find the global optima. Experimental results on synthetic and real datasets demonstrate the advantages of BMF over existing biclustering methods.

**Literature review**

A very fast algorithm for general matrix factorization of a data matrix present for use in the statistical analysis of high-dimensional data via latent factors. Such data are prevalent across many application areas and generate an ever-increasing demand for methods of dimension reduction in order to undertake the statistical analysis of interest. Our algorithm uses a gradient-based approach which can be used with an arbitrary loss function provided the latter is differentiable. The speed and effectiveness of our algorithm for dimension reduction is demonstrated in the context of supervised classification of some real high dimensional data sets from the bioinformatics literature. A simple update rule provide for computing the factorization and give supporting theoretical analysis. Finally, we perform a series of numerical experiments to show evidences of the good behavior of the numerical scheme. The ensemble NMF present for clustering biomedical documents in this paper. The performance of ensemble NMF was evaluated on numerous datasets generated from the TREC Genomics track dataset. With respect to most datasets, the experimental results have demonstrated that the ensemble NMF significantly outperforms classical clustering algorithms of bisecting K-means and hierarchical clustering. We compared four different methods for constructing an ensemble NMF. For clustering biomedical documents, this research is the first to compare ensemble NMF with typical classical clustering algorithms and validates ensemble NMF constructed from different graph-based ensemble algorithms. This is also the first work on ensemble NMF with Hybrid Bipartite Graph Formulation for clustering biomedical documents. It reported the advent age of microarray technology enables us to monitor an entire genome in a single chip using a systematic approach. Clustering, as a widely used data mining approach has been used to discover phenotypes from the raw expression data. However traditional clustering algorithms have limitations since they can not identify the substructures of samples and features hidden behind the data. Different from clustering, biclustering is a new methodology for discovering genes that are highly related to a
subset of samples. Several biclustering models/methods have been presented and used for tumor clinical diagnosis and pathological research. Experimental results on synthetic and real datasets demonstrate the advantages of BMF over existing biclustering methods. Besides the attractive clustering performance, BMF can generate sparse results (i.e., the number of genes/features involved in each biclustering structure is very small related to the total number of genes/features) that are in accordance with the common practice in molecular biology.\textsuperscript{16}

It reported an iterative normalized compression method for dimensionality reduction using Non-negative Matrix Factorization (NCMF). As a result, the basis matrix can be viewed as a compression matrix and the coefficient matrix becomes a mapping matrix. NCMF is simple, effective and only needs to initialize the mapping matrix. Experimental comparisons on text, biological and image data demonstrate that NCMF gains 21.02\% computational time reduction, 39.60\% sparsity improvement for mapping matrix and 8.59\% clustering accuracy improvement.\textsuperscript{17}

It introduces a novel approach to solve NMF problems, based on the use of an under approximation technique and show its effectiveness to obtain sparse solutions. This approach, based on Lagrangian relaxation, allows the resolution of NMF problems in a recursive fashion. We also prove that the under approximation problem is NP-hard for any fixed factorization rank, using a reduction of the maximum edge biclique problem in bipartite graphs. They test two variants of our under approximation approach on several standard image datasets and show that they provide sparse part-based representations with low reconstruction error. Our results are comparable and sometimes superior to those obtained by two standard sparse nonnegative matrix factorization techniques.\textsuperscript{17, 18}

It present two important ingredients are missing in the standard NMF methods, discriminant analysis with label information and geometric structure (manifold) in the data. Most of the existing variants of NMF incorporate one of these ingredients into the factorization. In this paper, we present a variation of NMF which is equipped with both these ingredients, such that the data manifold is respected and label information is incorporated into the NMF. To this end, we regularize NMF by intra-class and inter-class knearest neighbor (k-NN) graphs, leading to NMF-kNN, where we minimize the approximation error while contracting intra-class neighborhoods and expanding inter-class neighborhoods in the decomposition. They develop simple multiplicative updates for NMF-kNN and present monotonic convergence results. Experiments on several benchmark face and document datasets confirm the useful behavior of our proposed method in the task of feature extraction.\textsuperscript{18, 19}

A new unsupervised method proposes a new unsupervised method using Non-negative Matrix Factorization (NMF) to select sentences for automatic generic document summarization. The proposed method uses non-negative constraints, which are more similar to the human cognition process. As a result, the method selects more meaningful sentences for generic document summarization than those selected using LSA.\textsuperscript{20}

It has been reducing the size of the database using data mining techniques and NMF. In this research is presented three different algorithms such as Principal Component Analysis, Non-negative Matrix Factorization and Random Projection. Principal Component Analysis algorithm based on the different nature of the data set has the best results compared to other algorithms in the feature dimension reduction and data dis-play provides.\textsuperscript{21}

And many other publications discussed about BMF and its application on biclustering. Many biclustering algorithms have been proposed recently to explore the correlations between genes and samples and to identify the local gene-sample structures in microarray data.\textsuperscript{22}

The idea of biclustering is to characterize each sample by a subset of genes and to define each gene in a similar way. As a consequence, biclustering algorithms can select the groups of genes that show similar expression behaviors in a subset of samples that belong to some specific classes such as some tumor types, thus
identify the local structures of the microarray matrix data.\textsuperscript{22}
Several biclustering methods have been presented in the literature including BiMax, ISA (Iterative Signature Algorithm), SAMBA and OPSM (Order Preserving Submatrix). A systematic comparison and evaluation of these methods have been studied. Recently, Non-negative Matrix Factorization (NMF), as a useful tool for analysing datasets with non-negativity constraints, has been receiving a lot of attention.\textsuperscript{23-25} Nonnegative Matrix Factorization (NMF) factorizes an input nonnegative matrix into two nonnegative matrices of lower rank. In particular, NMF with the sum of squared error cost function is equivalent to a relaxed K-means clustering, the most widely used unsupervised learning algorithm.\textsuperscript{24,25} In addition, NMF with the I-divergence cost function is equivalent to probabilistic latent semantic indexing, another unsupervised learning method popularly used in text analysis.\textsuperscript{24, 25, 26} However, NMF can not produce the biclustering structures explicitly.

AIMS AND OBJECTIVES
In this paper, we extend standard NMF to Binary Matrix Factorization (BMF) for solving the biclustering problem: the input binary gene-sample matrix X is decomposed into two binary matrices W and H. The binary matrices W and H preserve the most important integer property of the input matrix and also explicitly designate the cluster memberships for genes and samples.\textsuperscript{26,27} As a result, BMF leads to a new biclustering model.

METHODODOLOGY
It was decided to adopt a case study approach for this paper as there is little existing research on measurement and identification of Environmental Projects Risks. It has been based on the descriptive research. This descriptive type research has been carried out using the questionnaire as the research tool for gathering the required data. Data gathering involved both reference material and a questionnaire survey. Sampling was simple random sampling and the data gathering instrument was the questionnaire. The author had already undertaken research in his field which had stimulated the measurement tools and the theoretical framework used to analyze this case study, based on BMF Method.

Sample
In selecting respondents, diversity was paramount so that opinions from a number of employment situations could be gathered. Consequently, different occupational groups in different sections were targeted. The employers included four public sectors in the company. The size of the lead and zinc Zanjan Company ranged from a low of 223 employees to a high of 410 with a mean size of 270 employees. Knowledge worker interviewees were in temporary or fixed-term employment contracts, representing independent contractors, short-term Environmental Projects employees hired by the organization and seasonal employees, also hired directly by the organization. Occupations included accountants, engineers, human resources specialists, information technology specialists (programmers, developers), quality assurance specialists, Environmental Projects managers, researchers, planners and resource conservation officers. In terms of gender, the contingent worker interviews included 27 men and 20 women and ages ranged from people in their early 20s to those who were near retirement. The education of respondents included seven individuals with no postsecondary degree, eight with a college or technical diploma, 21 with an undergraduate degree and 11 possessing a graduate degree. Managers were chosen from amongst a group who were responsible for the hiring and/or supervision of the selected contingent knowledge workers in each organization and represented the departments of the occupations listed above. Overall, 11 of the managers were male, while 12 were female and their ages ranged from 36 to 58s.

Data collection
Semi-structured, in-depth interviews were conducted with the contingent workers and their managers. Average length of interviews was 45 minutes in any month with half conducted face-to-face and half via telephone. Questions for the knowledge workers primarily explored their perceptions of the contingent employment relationship and included the more typical organizational measurements of
overall satisfaction with the employment arrangement, pay and hours of work. Secondly, due to various earlier discussions with contingent knowledge workers, research respondents were asked to rate how being in a contingent position affected their work/life balance, social distance (organizational socialization and integration) and knowledge sharing. Finally, because traditional empirical measurements of psychological contracts include notions of advancement, training and career development and more recent work has examined commitment, trust and work behavior, knowledge workers were asked to rate the effect of contingent work on their career goals, personal finances, promotion opportunities, autonomy on the job and organizational commitment. Overall, these measures were chosen so as to provide results relative to a number of different personal and job dimensions. A four-point forced choice scale was used for this aspect of the data collection. In total, all factors affecting the productivity of knowledge workers through interviews and questionnaires were extracted.

Data analysis

For the qualitative interview data, two researchers independently analyzed it to identify key issues and themes. For clustering of risks and projects are used the BMF method. Although NMF has shown its power in many applications, it cannot discover the biclustering structures explicitly. We extend the standard NMF to Binary Matrix Factorization (BMF), that is, elements of X are either 1 or 0 and we want to factorize X into two binary matrices W and H (thus conserving the most important integer property of the objective matrix X) satisfying $X \approx WH$. We will study both the theoretical and the practical aspects of BMF. Also, we present two algorithms of BMF for biclustering.

Penalty Function Method

In terms of nonlinear programming, the problem can be represented as:

$$
\min J(W, H) = \sum_{i,j} (X_{ij} - (WH)_{ij})^2 \\
\text{s.t. } H^2_{ij} - H_{ij} = 0 \\
W^2_{ij} - W_{ij} = 0
$$

Which can be solved by a penalty function algorithm and is programmed as follows

Algorithm 1: Penalty function method of BMF

Step 1 Initialize $\lambda$, $W$, $H$ and $\varepsilon$, Normalize $W$, $H$

Step 2 For $W$ and $H$, alternately solve:

$$
\min J(W, H) = \sum_{i,j} (X_{ij} - (WH)_{ij})^2 + \frac{1}{2} \lambda \left[ (H_{ij}^2 - H_{ij})^2 + (W_{ij}^2 - W_{ij})^2 \right]
$$

Step 3 if $(H_{ij}^2 - H_{ij})^2 + (W_{ij}^2 + W_{ij})^2 < \varepsilon$

$W = \Theta(W-0.5)$; $H = \Theta(H-0.5)$, break

Else $\lambda = : 10 \lambda$, return to 2.

Where the Heaviside step function is defined as

$$
\Theta(x) = \begin{cases} 
1 & x \geq 0 \\
0 & x < 0 
\end{cases}
$$

and $\Theta(\bullet)$ is element-wise operation: $\Theta(\bullet)$ is a matrix whose $(i, j)$th element is $\Theta(\bullet)_{ij}$.

Thresholding method

The second method is thresholding, in otherwords, finding the best thresholds $w$, $h$ for $W$ and $H$ respectively so that the minima of the following problem can be achieved:

$$
\min F(w, h) = \frac{1}{2} \sum_{i,j} (X_{ij} - (\Theta(W-w)\Theta(H-h))_{ij})^2
$$

Initial values of $W$, $H$ are given via the original NMF algorithm. As we can see, $\Theta(x)$ is non-smooth, so the problem is a non-smooth optimization problem. There are two implementations to conquer this difficulty.

Discretized method

We discretize the domain $\{(w, h) : 0 \leq w \leq \max(W), 0 \leq h \leq \max(H)\}$ and try on every grid point to search for optimal thresholds $(w^*, h^*)$.

Gradient Decent Method

We approximate the Heaviside function by the function

$$
\Theta(x) = \phi(x) = \frac{1}{1 + e^{-\lambda x}}, \lambda > 0 \quad \lambda > 0 \text{ is a large constant.}
$$
Then one can solve the replaced problem under the Gradient Decent method framework as Algorithm 2.

Algorithm 2 : Thresholding method of BMF

Step 1 Initialize \( w_0, h_0, k=0 \) Normalize \( W \) and \( H \)

Step 2 Compute gradient direction \( g_k \) of \( F(w, h) \). Select stepsize \( \alpha_k \)

Step 3 \( w_{k+1} = w_k - \alpha_k g_k, h_{k+1} = h_k - \alpha_k g_k \)

If some stop strategy is satisfied

\( W = : \Theta(W-w_{k+1}) ; H = : \Theta(H-h_{k+1}) \), break else \( k=k+1 \), turn to step 2

Comparison methods

We compare BMF-based algorithms with six other methods, BiMax, ISA, SAMBA, Binary Non-orthogonal Matrix Decomposition (BND), SNMF/R and nsNMF. The first three algorithms have been reported to be the best among the six biclustering methods. ISA and Bimax are implemented by the software BiCAT developed. SAMBA is implemented by EXPANDER. BND is implemented by PROXIMUS. SNMF/ and nsNMF are two recent NMF-based methods.

Reliability and validity

This method was tested using Cronbach’s alpha (its value was more than 89.8), it has been validated and confirmed by 87% of the experts, 88% of the managers and by company directors.

RESULTS AND DISCUSSION

Lead and zinc Zanjan Company is one of the largest producers in the Middle East.

Select the best threshold

In this section, we implement two algorithms (Penalty and Threshold methods) in Matlab software with various preprocessors. Results are shown in (Table 1)

Table 1 : Results of Penalty and Threshold methods

<table>
<thead>
<tr>
<th></th>
<th>1.05 , 0.07</th>
<th>1.05 , 0.05</th>
<th>1.05 , 0.04</th>
<th>1.06 , 0.07</th>
<th>1.07 , 0.07</th>
<th>1.08 , 0.07</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>Penalty</td>
<td>95.83</td>
<td>95.83</td>
<td>95.83</td>
<td>87.50</td>
<td>87.50</td>
</tr>
<tr>
<td></td>
<td>Threshold</td>
<td>91.67</td>
<td>87.50</td>
<td>87.50</td>
<td>91.67</td>
<td>87.50</td>
</tr>
<tr>
<td>NMI</td>
<td>Penalty</td>
<td>73.34</td>
<td>73.34</td>
<td>73.34</td>
<td>51.47</td>
<td>48.92</td>
</tr>
<tr>
<td></td>
<td>Threshold</td>
<td>58.90</td>
<td>51.47</td>
<td>51.47</td>
<td>58.09</td>
<td>51.47</td>
</tr>
</tbody>
</table>

With implementing of two above algorithms in various thresholds and in finally for matrix discretization, then we select final upper threshold and lower threshold (1.05 and 0.07), because of ACC and NMI have best answers for this range.

BMF-Penalty method

Results of Penalty method in matlab 2010 has shown Fig. 1.

BMF- Threshold method

Result of Threshold method in matlab 2010 has shown Fig. 2.
With the implementation of the two algorithms (Penalty and Threshold algorithms), NMI and ACC indexes for Penalty algorithms perform better than the threshold algorithm on the data being studied. The results of threshold algorithms for low-density matrices are better than penalty algorithm. Penalty algorithms for high-density matrices are better results. Thus, based on the relatively high dispersion of the studied matrices, the results of Penalty algorithm are better than Threshold algorithms. With the implementation of the nsNMF and SNMF / R algorithm in MATLAB 2010, SAMBA in XPANDER6 software and ISA and Bimax algorithms in BicAT_v2.22 software, results are shown in Table 2 and Fig. 3. Finally, the accuracy of the results in Table 2, BMF algorithms are better performance than to other algorithms.

**Table 2 : Results of ACC and NMI for biclustering algorithms**

<table>
<thead>
<tr>
<th></th>
<th>Model-base</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BMF Penalty</td>
</tr>
<tr>
<td>ACC</td>
<td>95.83%</td>
</tr>
<tr>
<td>NMI</td>
<td>73.34%</td>
</tr>
</tbody>
</table>

**Fig. 2 : Results of Threshold method**

**Fig. 3 : Comparison of biclustering algorithms results based on ACC and NMI**

NMI=0.5890, ACC=0.9167, a=0.57, b=0.53, k=2, cutoff X =1.05, cutoff Y=0.07
The results of penalty algorithms for NMI and ACC indexes are better than the other algorithms and threshold algorithm is in the next ranking.

**Dispersion index**

In this section, the calculation of retreats matrices from the initial decomposition matrix, total number of the non-zero numbers divide to total number of the zero numbers in each matrix, Multiply this numbers by 100, it is shown in Table 3.

This method was tested using Cronbach’s alpha (its value was more than 89.8), it has been validated and confirmed by 87% of the experts, 88% of the managers and by company directors.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>W (%)</th>
<th>H (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMF-penalty</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>BMF-threshold</td>
<td>44.23%</td>
<td>50%</td>
</tr>
<tr>
<td>SNMF/R</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>nsNMF</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>

In contrast, many of the contract and short-term Environmental Projects contingent workers were able to justify their employment uncertainty, at least in the short-term because of the variety, development, learning and other opportunities that their various experiences provided. But they were also keenly aware of the negative impact of contingent employment on their careers, some thought that they had little ability to control or influence the length of their contracts and, as a result, had little control in other aspects of their lives. The most severe effects, as revealed in both the interview and scale data, were experienced in career goals, opportunities for training and development and chances for promotion. Employers need to be more attuned to what kinds of contingent work relationships they are involved with and how they may change over time. The example of the seasonal workers underscores the need for contingent work force strategies that take into account the longer term. Hence, employers who are serious about talent management may want to look more closely at how they might contribute to the development of contingent workers and retain the intellectual capital that such employees bring to or acquire while they are inside the organization. The results also indicate that employers have considerable room to improve the integration of contingent and non-contingent workers. Employers could also do more to capitalize on the knowledge and experience of contingent workers. The loss of good people, as well as the loss of knowledge and expertise when the organization was not able to offer ongoing work or regular positions must be considered. Previous research indicates that contingent workers provide a valuable source for enhancing the intellectual capital of firms by importing new knowledge and techniques.26

With few exceptions, knowledge workers in this study made it clear that they believed the sharing and transfer of their knowledge and expertise was a personal and professional responsibility.

**CONCLUSION**

We have defined and classified the effective elements of Environment Project risks and clustered them using BMF. Consequent to this analysis, we have presented clusters for improving project risks, which were verified and validated in a case study of Environmental Projects. In this paper, we propose BMF to identify the biclustering structures in microarray data. In fact, several papers have discussed about the biclustering aspect of NMF. But the key difficulty is that one can not identify the binary relationships
between genes and samples exactly since the resulting matrices W and H are not binary. This can be solved via BMF. In addition, since W and H are binary, BMF offers a framework for simultaneously clustering the genes and samples. The framework is able to perform implicit feature selection and provide adaptive metrics for biclustering. All of these properties are preferable for clustering in high-dimensional data. As for future work, we will investigate more sophisticated discretization methods on real data to improve numerical performance. Our current discretization method is time-consuming, especially for data having complex structures. We will develop more efficient strategies for discretizing original microarray data. One possible follow-up is the comparison of the proposed method with other models, such as the fuzzy topics and neuron fuzzy methods.

ACKNOWLEDGEMENT

This work was supported by a grant from Islamic Azad University, Zanjan, Iran.

REFERENCES

16. Mikkel N. Schmidt Ole Winther and Lars Kai Hansen, Bayesian non-negative


Instructions to authors

It is a condition for publication that the authors must give an undertaking in the writing at the time of submission of papers that the manuscripts (research papers) submitted to JERAD (Journal of Environmental Research And Development) have not been published and have not been submitted for publishing elsewhere, manuscripts are their original work.