MUTUAL INFORMATION OF TENSORS FOR MEASURING THE NONLINEAR CORRELATIONS

Liangfu Lu[1], *Xiaohan Ren*[1], *Yongheng Jia*[2], *Rize Jin*[3]*, *Zhenghai Huang*[1]

[1]School of Mathematics, Tianjin University, Tianjin, China;
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China;
[3]School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin, China;
*Corresponding Author: jinrize@tjpu.edu.cn

ABSTRACT

Correlation analysis has been proposed to measure the relationship among different variables, with application in multi-view dimension reduction. However, the existing methods usually are used by covariance in a linear way rather than the nonlinear effects being considered among multiple variables and only few works on nonlinear interaction of two variables have been considered. In this paper we propose a nonlinear analysis of multiple(more than two)variables based on mutual information for tensors analysis(MITA) firstly. In addition, we extend the mutual information matrix analysis directly to mutual information tensor analysis and show the mutual information formula for multiple variables theoretically. Experiments on multi-view dimension reduction, including biometric structure prediction and internet advertisement classification problems, have been done to illustrate the effectiveness of the proposed method, especially in the case of low dimensional subspace.

***Index Terms*—** Mutual information, tensor, nonlinear correlation, dimension reduction

1. INTRODUCTION

In real-world applications, solutions of many problems are influenced by multiple variables or multiple factors. Therefore, finding methods for multivariate correlation analysis become important issue in the field of machine learning, high-order statistical analysis and image processing. For example, it can be performed by building a small set of new features through linear or non-linear transformation of the original data, i.e., by performing a linear or nonlinear feature extraction(FE)[1]. In an urban transportation system, the influencing factors of traffic flow are analyzed by the multivariate clustering method [2]. In addition, the similarity matrix between variables is used in image registration[3]. Since multivariate correlation analysis plays an important role in these applications, it can clarify the correlation (interaction) between variables or factors and make possible for us to reduce the dimension of the data and minimize the redundant information by measuring the interactions among factors.

In statistics, Canonical Correlation Analysis (CCA) is a kind of method to effectively analyze the correlation between two variables. When the variable is a vector, the correlation between the two variables is usually measured by a similarity matrix or a covariance matrix (two-dimensional array). While the correlation of the two variables constitutes an N-dimensional array, which is covariance tensor, when the variable is a matrix or tensor. Kim T K et al. [4-5] proposed a method termed tensor canonical correlation analysis (TCCA), which is a generalization of canonical correlation analysis to multidimensional arrays (tensor) and apply this for action/gesture classification in videos. Hsieh C Y et al. [6-7] considered each expanded matrix in a tensor as an individual to estimate the similarity between two tensors, which can effectively utilize the characteristics of each expanded matrix. Luo Y et al. [8] presented that the main problem of multi-view specification correlation maximization is equivalent to finding the best rank-1 approximation of the data covariance tensor, which can be effectively solved using the alternating least squares (ALS) algorithm. However, whether a variable is a vector, a matrix or even a tensor, it is essentially a linear correlation analysis of two variables or multiple variables. Because covariance, correlation coefficient and similarity are all linear relationships between variables. Zhao X et al. [9] proposed a new mutual-information matrix analysis to study the nonlinear interactions of multivariate time series, but in fact they considered still two variables. To our best knowledge, the nonlinear correlation among multiple variables haven't been considered by the current approaches. To solve this problem, we propose a nonlinear analysis of multivariate interactions based on mutual information tensors analysis(MITA) in this paper.

The rest of the paper is organized as follows. In Section 2 we present the basic preliminaries of tensor algebra, some information about previous work in mutual information matrix. We propose our new method for multivariate nonlinear correlation analysis in Section 3. The experimental results are shown in Section 4 and conclusions and future work of the paper are concluded in Section 5.

2. PRELIMINARIES

In this section, we mainly present the definitions of basic operations on tensors[10] and the related nonlinear correlation analysis of the two variables based on the mutual information, especially the matrix conditions [9]. In the whole paper, the scalars will be denoted by italic letters (e.g., x), vectors by boldface lowercase letters (e.g., $\mathbf{x}$), matrices by boldface capital letters (e.g., $\mathbf{X}$) and tensors by boldface Euclid math one letters (e.g., $\boldsymbol{\mathcal{A}}$).

2.1. Tensor notation, operations and decompositions

In general, a *m*-order tensor $\boldsymbol{\mathcal{A}} = \left(x_{i_1 \cdots i_m}\right) \in \mathbf{R}^{N_1 \times N_2 \times \cdots N_m}$ is a multi-array, where $i_j = 1, \ldots, N_j$ for $j = 1, \ldots, m$ .The mode-*n* matricization of a tensor $\boldsymbol{\mathcal{A}} \in \mathbf{R}^{N_1 \times N_2 \times \cdots N_m}$ is denoted by $\mathbf{X}_{(n)}$ and arranges the mode-*n* fibers to be the columns of the resulting matrix. The tensor *n*-mode product, i.e., multiplying a tensor by a matrix (or a vector) in mode *n*.

The *n*-mode matrix product of a tensor $\boldsymbol{\mathcal{A}} \in \mathbf{R}^{N_1 \times N_2 \times \cdots N_m}$ with a matrix $\mathbf{U} \in \mathbf{R}^{J \times N_n}$ is denoted by $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{A}} \times_n \mathbf{U}$ and $\boldsymbol{\mathcal{B}} \in \mathbf{R}^{N_1 \times N_2 \times \cdots N_{n-1} \times J \times N_{n+1} \times \cdots N_m}$, each element in $\boldsymbol{\mathcal{B}}$ is

$$b_{i_1 i_2 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum_{i_n=1}^{N_n} x_{i_1 i_2 \cdots i_N} u_{j i_n} \tag{1}$$

The product of $\boldsymbol{\mathcal{A}}$ with a sequence of matrices $\left\{\mathbf{U}_P \in \mathbf{R}^{J_p \times N_p}\right\}_{p=1}^{m}$ is a $J_1 \times J_2 \times \cdots J_m$ tensor denoted by

$$\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{A}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_m \mathbf{U}_m \tag{2}$$

The *n*-mode vector product of a tensor $\boldsymbol{\mathcal{A}} \in \mathbf{R}^{N_1 \times N_2 \times \cdots N_m}$ with a vector $\mathbf{v} \in \mathbf{R}^{N_n}$ is denoted by $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{A}} \bar{\times}_n \mathbf{v}$ and $\boldsymbol{\mathcal{B}} = \left(b_{i_1 i_2 \cdots i_{n-1} i_{n+1} \cdots i_N}\right) \in \mathbf{R}^{N_1 \times N_2 \times \cdots N_{n-1} \times N_{n+1} \times \cdots N_m}$, each element in $\boldsymbol{\mathcal{B}}$ is

$$b_{i_1 i_2 \cdots i_{n-1} i_{n+1} \cdots i_N} = \sum_{i_n=1}^{N_n} x_{i_1 i_2 \cdots i_N} v_{i_n} \tag{3}$$

The product of $\boldsymbol{\mathcal{A}}$ with a sequence of vectors $\left\{\mathbf{v}_P \in \mathbf{R}^{N_p}\right\}_{p=1}^{m}$ is a scalar denoted by

$$b = \boldsymbol{\mathcal{A}} \bar{\times}_1 \mathbf{v}_1 \bar{\times}_2 \mathbf{v}_2 \bar{\times} \cdots \bar{\times}_m \mathbf{v}_m \tag{4}$$

2.2. Mutual-information matrix analysis

Let $\mathbf{X} = \left\{\mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_n\right\} \in \mathbf{R}^{m \times n}$ and $\mathbf{x}_i \in \mathbf{R}^m$, the mutual information $I\left(\mathbf{x}_i, \mathbf{x}_j\right)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined by

$$\begin{aligned} I_{ij} &\triangleq I\left(\mathbf{x}_i, \mathbf{x}_j\right) = H\left(\mathbf{x}_i\right) + H\left(\mathbf{x}_j\right) - H\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ &= \sum p\left(\mathbf{x}_i, \mathbf{x}_j\right) \log \frac{p\left(\mathbf{x}_i, \mathbf{x}_j\right)}{p\left(\mathbf{x}_i\right) p\left(\mathbf{x}_j\right)} \end{aligned} \tag{5}$$

Where H denotes the Shannon entropy. As an effective information-theoretic tool, the mutual information expresses the shared information between $\mathbf{x}_i$ and $\mathbf{x}_j$. Moreover, it estimates how much knowing one of two variables reduces the uncertainty about the other as well, which takes both the linear and nonlinear dependence into consideration. A normalized mutual information as follows:

$$\hat{I}_{ij} = \frac{2 I_{ij}}{H\left(x_i\right) + H\left(x_j\right)} \tag{6}$$

The normal mutual information between each pair of two variables construct a mutual-information matrix :

$$\mathbf{I} = \begin{bmatrix} \hat{I}_{11}, \hat{I}_{12}, \cdots, \hat{I}_{1n} \\ \hat{I}_{21}, \hat{I}_{22}, \cdots, \hat{I}_{2n} \\ \vdots \quad \vdots \quad \vdots \\ \hat{I}_{n1}, \hat{I}_{n2}, \cdots, \hat{I}_{nn} \end{bmatrix} \tag{7}$$

Where $\hat{I}_{ij} \equiv \hat{I}_{ji}$ for $\forall i, j \in [1, n]$,so the matrix $\mathbf{I}$ is also symmetric.

A measure of *n*-dimensional mutual information to analyze the interactions among multivariate, which based on the heterogeneity of the eigenvalues of the mutual information matrix is given by[9]:

$$I = 1 + \frac{\sum_{i=1}^{n} \frac{|\lambda_i|}{\sum_{j=1}^{n} |\lambda_j|} \log\left(\frac{|\lambda_i|}{\sum_{j=1}^{n} |\lambda_j|}\right)}{\log(n)} \tag{8}$$

Where λ_i is the *i*-th eigenvalue of $\mathbf{I}$ and $0 \le I \le 1$.

3. MUTUAL-INFORMATION TENSOR ANALYSIS

We consider the mutual information of multiple variables directly instead of measuring the interaction of multiple variables through the mutual information matrix of two variables. The definition of mutual information of three variables based on Venn diagram is given by[11]:

$$\begin{aligned} I\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\right) &= H\left(\mathbf{x}_1\right) + H\left(\mathbf{x}_2\right) + H\left(\mathbf{x}_3\right) - H\left(\mathbf{x}_1, \mathbf{x}_2\right) \\ &\quad - H\left(\mathbf{x}_1, \mathbf{x}_3\right) - H\left(\mathbf{x}_2, \mathbf{x}_3\right) + H\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\right) \end{aligned} \tag{9}$$

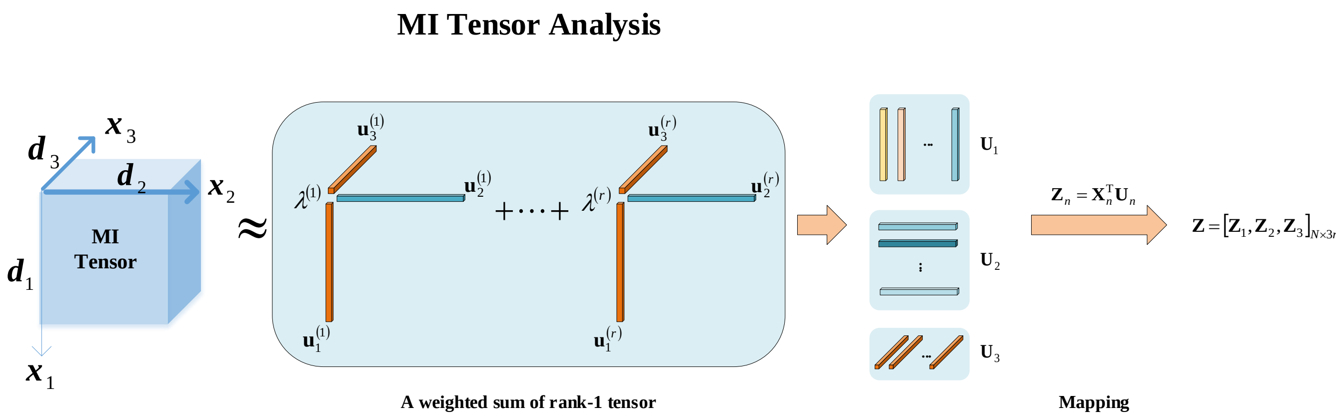

Fig. 2. Multi-views dimension reduction based on MITA. First, the MI tensor of multiple views is computed by using feature matrices, which represents the correlation information between all views. Then the MI tensor is approximated as a weighted sum of rank-1 tensor, we use the transformation matrix $\mathbf{U}_p$ for the p-th view to map the original $\mathbf{X}_p$ to the low dimensional $\mathbf{Z}_p$ in the common subspace,. Finally, we use the concatenation of $\mathbf{Z}_p$.

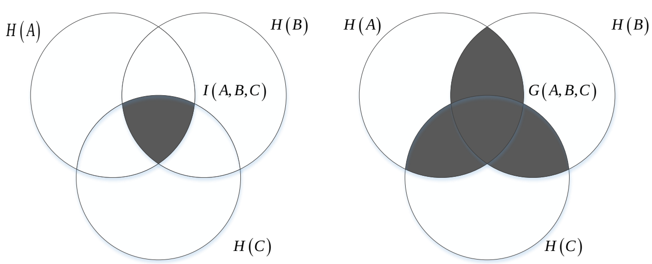

Fig. 1. three-dimensional MI (black) and G-MI(black) based on Venn diagram.

We can define the mutual information of n variables based on Wayne diagram, It represents shared information between n variables. It denotes by(10). A normalized mutual information(N-MI) denotes by(11).

$$I(\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_n)=\sum H(\mathbf{x}_i)-\sum H(\mathbf{x}_i,\mathbf{x}_j)+\sum H(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)-\cdots(-1)^{n-1}H(\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_n) \quad (10)$$

$$\hat{I}(\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_n)=\frac{nI(\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_n)}{\sum H(\mathbf{x}_i)} \quad (11)$$

With the increase of n, the calculation of MI is very complicated and may be negative, it may make no sense to select multiple variables. To avoid these situations, we can use the high-dimensional generalized redundancy given by[12], in this paper, we call it generalized mutual information(G-MI). It represents the shared information of all permutations and combinations of n variables. Generalized mutual information contains more information than mutual information.

$$G(\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_n)=H(\mathbf{x}_1)+\cdots+H(\mathbf{x}_n)-H(\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_n) \quad (12)$$

Without losing generality, we take mutual information and generalized mutual information of three variables as examples. Fig.1 shows the mutual information and generalized mutual information based on Wayne diagram.

In contrast to mutual-information matrix analysis, where only considered the pairwise mutual information, we propose mutual-information tensor analysis for multiple variables, i.e. MITA, G-MITA, N-MITA and can be applied to dimension reduction from multiple views. The difference between the three methods is the mutual information used to construct tensors. Extracting different types of features to represent instances in different views leads to multiple feature matrices $\left\{\mathbf{X}_P=\left[\mathbf{x}_{p1},\mathbf{x}_{p2},\cdots\mathbf{x}_{pd_p}\right]^{\mathrm{T}}\in\mathbf{R}^{d_p\times N}\right\}_{p=1}^{n}$.

Without losing generality, n is set at 3. The diagram of the multi-view dimension reduction method using the MITA is shown in Fig.2. The same is true for G-MITA and N-MITA. The different sets of features are used to calculate the data MI (N-MI/G-MI) tensor $\mathcal{I}_{123}\in\mathbf{R}^{d_1\times d_2\times d_3}$, which is subsequently decomposed as a weighted sum of rank-1 tensors, i.e., CP decomposition[13].

$$\mathcal{I}_{123}\approx[\lambda;\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3]\equiv\sum_{k=1}^{r}\lambda^{(k)}\mathbf{u}_1^{(k)}\circ\mathbf{u}_2^{(k)}\circ\mathbf{u}_3^{(k)} \quad (13)$$

Where $\boldsymbol{\lambda} \in \mathbf{R}^{r}$, $\circ$ is the outer product, $r \le \min\left(d_1, d_2, d_3\right)$ is the reduced dimension. $\mathbf{U}_n \in \mathbf{R}^{d_n \times r}$ is a transformation matrix from $\mathbf{X}_n$ to $\mathbf{Z}_n$ by $\mathbf{Z}_n = \mathbf{X}_n^{\mathrm{T}} \mathbf{U}_n$.

The mutual information tensor among all views is calculated as

$$\mathcal{I}_{12\cdots n} = \left(\mathcal{I}_{i_1 i_2 \cdots i_n}\right) = \left(I\left(\mathbf{x}_{1i_1}, \mathbf{x}_{2i_2}, \cdots \mathbf{x}_{ni_n}\right)\right) \tag{14}$$

Where $I\left(\mathbf{x}_{1i_1}, \mathbf{x}_{2i_2}, \cdots \mathbf{x}_{ni_n}\right)$ is mutual information and $\mathcal{I}_{12\cdots n} \in \mathbf{R}^{d_1 \times d_2 \times \cdots d_n}$.

The maximize the mutual information I of MI tensor is equivalent to finding the best rank-1 approximation of the tensor $\mathcal{I}_{12\cdots n}$ [8], it denotes by

$$\begin{gathered} \arg\max_{\mathbf{u}_p} I = \mathcal{I}_{12\cdots n} \bar{\times}_1 \mathbf{u}_1^{\mathrm{T}} \bar{\times}_2 \mathbf{u}_2^{\mathrm{T}} \cdots \bar{\times}_n \mathbf{u}_n^{\mathrm{T}} \\ s.t.\ \mathbf{u}_p^{\mathrm{T}} \mathbf{u}_p^{\mathrm{T}} = 1,\ p = 1, 2 \cdots n \end{gathered} \tag{15}$$

4. EXPERIMENTS

In this section, we verify the validity of the proposed mutual information tensor for biometric feature structure prediction and advertisement classification following [8]. The first set of experiments is conducted in the transformation settings, that is, the test data are available in the training stage, and the training set and test set are both used to map the samples into a low-dimensional subspace. Furthermore, considering another situation, the test data may not be available in training phrase, for example some data are generated continuously, the cost of retraining after obtaining new data is too high. Therefore, in the second group of experiments, we randomly selected some data as the training set, and the rest of the data are used in the test phase. Only training data is used to find the subspace in the second set experiments. In the two sets of experiments, we use regularized least squares (RLS) as the base learner following [8]. Twenty percent of the test data are used for validation. We use the implementation of RLS in [14] as learners in our experiment. The regularization parameter of RLS is tuned over $\left\{2^i \,\middle|\, i = -15, \ldots, 16\right\}$, the regularization parameter performs best are used to test the performance of different methods. The setting of RLS learners in all experiments are the same. We use the classification accuracy as the evaluation criterion.

4.1. Biometric Structure Prediction

We use the dataset SecStr[15] in this set of experiments, which is a benchmark dataset for evaluating semi-supervised systems. The dataset contains 84K instances, we randomly

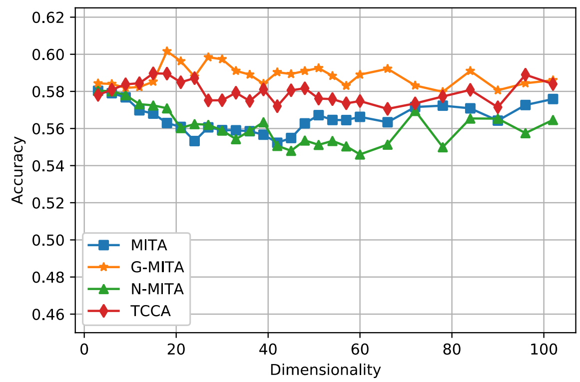

Fig. 3. Accuracies of Different Methods at Different Dimensions of Subspace on SecStr Dataset

Table 1. Accuracies of Different Methods at Their Best Dimensions on SecStr Dataset

Methods	Accuracies(%)
TCCA	58.97±1.21
MITA	58.03±2.06
G-MITA	**60.16±0.87**
N-MITA	58.04±1.32

select 100 instances as labeled samples. TCCA proposed in [8], which generalize CCA to handle multi-view data in a direct and natural way, has achieved promising results in multi-view dimension reduction, so in our paper, we compare our methods with it.

There are 15 categorical attributes in the provided features, each of which is represented by a 21-dimensional sparse binary vector. Following [8], we divided these features into three views:

- View-1: 5 categorical attributes based on the left context.
- View-2: 5 categorical attributes based on the middle context.
- View-3: 5 categorical attributes based on the right context.

The dimension of each view is 105(5×21).

We compare the performance of MITA, G-MITA and N-MITA with TCCA. For all of the four methods, we run experiments 50 times for each dimension r in $\left\{3, 6, 9 \cdots 60, 66, \cdots 102\right\}$. Fig.3 shows the performance of the compared methods in different dimensions of the common subspace. From the result, Our MITA performs slightly better than N-MITA, N-MITA does not perform well in this situation. This may be caused by that the mutual information has already measured the linear or nonlinear relation between different variables, it makes no sense to further

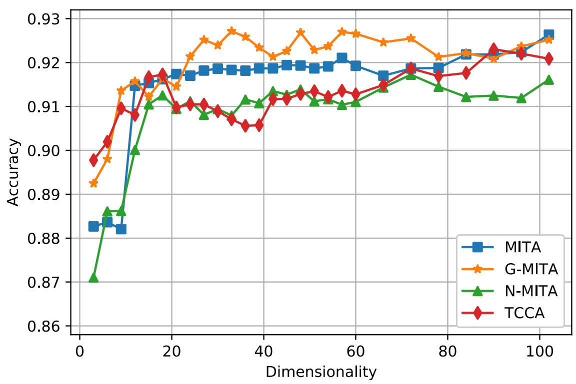

Fig. 4. Accuracies of Different Methods at Different Dimensions of Subspace on Internet Advertisements Data Set

normalize it. TCCA does perform better than MITA and N-MITA, but our G-MITA performs the best over other three methods. G-MITA consider not only the mutual information between multiple variables, but also the mutual information between each pair of variables. This shows that our method to use mutual information to measure the correlation among multiple variables is effective. For the 4 methods, the performance at their best dimension is showed in Table 1. From the result, we can see that G-MITA has the best accuracy. What's more, the G-MITA also has the least standard deviation, which means it is more "stable" than other methods.

4.2. Advertisement Classification

In this set of experiments, we use the Internet Advertisements Data Set[16], which contains 3279 instances and each instance has 1558 attributes. This dataset represents a set of possible advertisements on Internet pages, the task is to predict whether an image is an advertisement. We consider the situation that the test data maybe not available in training phrase. Therefore, in this group of experiments, we randomly selected 200 samples as the training set, and the rest of the dataset are used in the test phase, only the training set is used to find the subspace.

As in [8], We use the features as described in [16] and omit the attributes that have missing value. The remained 1555 binary attributes were divided into three views as follows:

- View-1: features from URL terms, alt terms and caption terms, 588 dimensions.
- View-2: features from URL of the current site, 495 dimensions
- View-3: features from anchor URL terms, 472 dimensions

As in the first set of experiments, we run experiment for

Table 2. Accuracies of Different Methods at Their Best Dimensions on Internet Advertisements Data Set

Methods	Accuracies(%)
TCCA	92.30 ± 0.31
MITA	92.63 ± 0.25
G-MITA	$\mathbf{92.72 \pm 0.30}$
N-MITA	91.72 ± 0.53

50 times in each dimension r in $\{3, 6, 9 \cdots 60, 66, \cdots 102\}$ and average the result. Fig.4 shows the performance of the compared methods. From Fig.4, we can see that the classification accuracy increase with the dimension r. Our MITA and G-MITA both perform well in this situation and G-MITA performs best out of four methods. N-MITA still does not perform well in this situation, the reason may be the same as we mentioned before. TCCA is slightly better than N-MITA but still not comparable with MITA and G-MITA. For the 4 methods, the performance at their best dimension is showed in Table 2. From the result, we can see that MITA and G-MITA both perform better than TCCA, G-MITA still performs best.

5. CONCLUSION

Tensor canonical correlation analysis takes into account the high-order statistics (correlation information) between all feature views, which can be applied to multi-view dimensionality reduction. While the related information only measures the linear correlations among multiple variables and ignores the nonlinear interaction of multiple ones. To address this problem, we have proposed a nonlinear analysis of multiple variables interactions which is termed mutual information of tensors.

After the experimental verification on several tasks of multi-view dimensionality reduction, the results demonstrate that: 1) Mutual information performs better than covariance in measuring the nonlinear interaction of variables. 2) Using mutual information directly is better than using normalized mutual information to measure the interaction of multiple variables. 3) By extending mutual information to multiple variables, the proposed G-MITA method is superior to other methods, especially in the case of low dimensional commons subspace. 4) G-MITA is not only performs best among these methods, but also has a more "stable" performance on those tasks.

6. REFERENCES

[1] Jukic, A and M. Filipovic, "Supervised feature extraction for tensor objects based on maximization of

mutual information." *Pattern Recognition Letters*, vol. 34, no. 13, pp. 1476-1484, Oct. 2013.

[2] Celikoglu, H. B. , and M. A. Silgu . “Extension of Traffic Flow Pattern Dynamic Classification by a Macroscopic Model Using Multivariate Clustering.” *Transportation Science*, vol. 50, no. 3, pp. 966-981, Aug. 2016.

[3] Xiongmei, Zhang , et al. "SAR Image Registration Method Based on Improved SIFT." *Computer Engineering*, vol. 41, no. 1, pp.223-226, 2015.

[4] Kim, T. K. , S. F. Wong , and R. Cipolla . "Tensor Canonical Correlation Analysis for Action Classification." *2007 IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, pp. 1352-1359, 2007.

[5] Kim, T. K. , and R. Cipolla . "Canonical Correlation Analysis of Video Volume Tensors for Action Categorization and Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no.8, pp. 1415-1428, Aug. 2009.

[6] Hsieh, C. Y. , and W. Y. Lin . "Effective fusing the factored matrices in dual tensors for action recognition." *IAPR International Conference on Machine Vision Applications IEEE*, pp. 386-389, 2015.

[7] Hsieh, C. Y. , and W. Y. Lin . "Video-based human action and hand gesture recognition by fusing factored matrices of dual tensors." *Multimedia Tools and Applications*, vol.76, no. 6, pp. 7575-7594, Mar. 2017.

[8] Luo, Y, et al. "Tensor Canonical Correlation Analysis for Multi-view Dimension Reduction." *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111-3124, Nov. 2015.

[9] Zhao, X. J. , P. Shang , and J. Huang . "Mutual-information matrix analysis for nonlinear interactions of multivariate time series." *Nonlinear Dynamics*, vol. 88, no. 1,pp. 477-487, Apr. 2017.

[10] Kolda, Tamara G. , and T. Gibson . "Multilinear operators for higher-order decompositions." *Office of Scientific & Technical Information Technical Reports*, 2006.

[11] Wang, Bo , and Y. Shen . "A method on calculating high-dimensional mutual information and its application to registration of multiple ultrasound images. " *Ultrasonics*, vol. 44, no. 4, pp. e79-e83, 2006.

[12] Kraskov, Alexander, H. Stögbauer, and P. Grassberger. "Estimating mutual information." *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol. 69, no. 6 Pt 2, 2006.

[13] Kolda, Tamara G, and B. W. Bader. "Tensor Decompositions and Applications." *Siam Review*, vol. 51, no. 3,pp.455-500, Sep. 2009.

[14] Pahikkala T, Airola A, “RLScore: regularized least-squares learners.” *The Journal of Machine Learning Research*, vol. 17, no. 1,pp. 7803-7807, Jan. 2016.

[15] http://www.kyb.tuebingen.mpg.de/ssl-book

[16] Kushmerick, N, "Learning to remove Internet advertisements." *Conference on Autonomous Agent ACM*, pp. 175-181, 1999.