Supervised Tensor Learning with Applications

BY

NESHAT MOHAMMADI
B.Sc., Shahid Beheshti University, 2009

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:

Pillip S. Yu, Chair and Advisor
Besma Smida
Ashkan Sharabinai, Mechanical and Industrial Engineering

to my beloved Parents

ACKNOWLEDGMENTS

I would like to express my appreciation and sincere gratitude to many people for their continuing support leading to this thesis. First of all, I would like to express my gratitude to my adviser, Professor Philip S. Yu, for his patience, guidance, supervision and trust throughout this work. His supportive guidance helped me overcome the problems and his trust gave me the courage to challenge the obstacles and enjoy exploring the field. I would also like to thank Mojtaba Soltanalian for supporting my research and sharing his lab's equipment with me.I take this opportunity to thank my fellow colleague Lifang He for her helps and encouragement through out my masters at University of Illinois at Chicago. Her knowledge in different aspects of computer science and machine learning was quite helpful when dealing with corner-cases in my research. I also thank Jon Mann for his valuable comments during writing my thesis. Many thanks to Besma Smida and Ashkan Sharabiani for serving on my thesis committee and providing their valuable insights on my research. I would especially like to thank my beloved parents for their constant caring, encouragement, understanding and sacrifice throughout these years. I deeply appreciate their spiritual and financial support and would like to dedicate this work to them with all my love. Last but not the least, many thanks to my little sister Fateme for her endless love and support.

NM

PREFACE

This dissertation is an original intellectual product of N. Mohammadi. All of the work presented here was conducted at the University of Illinois at Chicago and under supervision of Professor Philip S. Yu's.

Neshat mohammadi
July 21, 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DTI	Diffusion Tensor Imaging
fMRI	Functional Magnetic Resonance Imaging
DuSK	A Dual Structure-preserving Kernel
NPL	Natural Language Problem
PCA	Principle Component Analysis
RBF	Radial Based Function
SSS	Small Sample Size
SHTM	Support Higher- Order Tensor Machine
SNTF	Supervised Non-negative Tensor Factorization
STL	Supervised Tensor Learning
STM	Support Tensor Machine
SVM	Support Vector Machine

SUMMARY

In this thesis, a new supervised tensor learning (STL) approach with application to neuroimages has been studied and implemented. We applied our proposed polynomial kernel-based approach in order to analyze HIV infections based on fMRI and DTI brain images. The goal of this project was to achieve a more accurate prediction for HIV diagnosis using fMRI and DTI images of the brain.

To achieve this goal we tried to improve the accuracy of the STL model by directly using tensor data as an input. Then, in order to solve STL problems, a structure-preserving feature mapping in addition to CP decomposed results has been defined to derive a Dual Structure-preserving Kernel (DuSK) in the tensor product feature space. Broadly, DuSK is a general framework to convert any vector-based kernel function to an equivalent tensorial representation.

Different from traditional STL frameworks, that usually intend to use linear models, our approach was based on a nonlinear kernel method and tensor factorization techniques that can preserve the multi-way structures of tensorial data. We investigated the performance of DuSK together with Support Vector Machine (SVM) for HIV infection classification on tensorial fMRI and matricized DTI data sets.

The experimental results are presented with details in evaluation chapter. According to our experiments, that DuSK with a nonlinear kernel can effectively boost classification performance in HIV data sets, and the choice of optimal kernel depends on the nature of the input data. Specifically, DuSK with an RBF kernel performs better on fMRI data, while DuSK with a polynomial kernel is better for DTI data.

CHAPTER 1

INTRODUCTION

1.1 Motivation

With advancement in medical imaging, the importance of using medical imaging specially neuroimages to attain a more accurate diagnosis for disease came to notice. Currently there are two principal topics of concern in the fields of pattern recognition, computer vision, and image processing: data representation and classifier design.

Tensorial representation of data is very common in representing real world data sets specially in the medical imaging. The problem is most of the traditional classifiers only accept vectors or matrices as an input. Although tensor can be reshaped to vector and matrice to comply the input requirement of the classifier, several studies have presented that this direct reshaping breaks the natural structure and correlation in the original data and contribute to curse of dimensionality and Small Sample Size (SSS) problems. Current studies focus on two problems: (Kolda and Bader, 2009)

- **1) How to formulate other learning problems as tensor decomposition**
- **2) How to compute tensor decompositions under weaker assumptions**

Within the past 20 years, researchers have given strong attention to data representation methods such as tensor decomposition and multilinear subspace learning in order to address the curse of dimensionality and Small Sample Size problem (SSS) problems. Currently, in order to generalize the Support Vector Machine (SVM) learning framework to tensor patterns, researchers have developed several method such

as Support Tensor Machine (STM) (Tao et al., 2005), Support Higher-Order Tensor Machine (SHTM) (Hao et al., 2013) and other methods which are the extension of conventional classifier methods that can accept tensor as an input.

There has been increasing interest in the mathematics of higher-order tensors and development of efficient learning machines for tensor classification. Hence, tensor representation is becoming widely used to formulate optimization problems in different fields of study such as quantum information theory, signal processing, machine learning, algebraic statistics, and others, specially, medical imaging . Tensor representations provide a concise mathematical framework for formulating and solving many problems in various areas. In addition, tensor representation is useful to reduce the Small Sample Size problem (SSS) in discriminative subspace selection (Tao et al., 2005).

Various machine learning classification methods have been used for medical diagnosis. In particular, SVMs are supervised learning methods that have been widely and successfully used for medical diagnosis (He et al., 2014)(Hardoon and Shawe-Taylor, 2010). Among the existing SVM based methods, the input data are mostly represented as vectors. However, this discards the structure information in the Neuroimaging data and may lead to performance degradation.

In this thesis, we focus on the usage of SVM method for tensor Neuroimaging data analysis, in conjunction with tensor kernel methods. In this study problems of Supervised Tensor Learning (STL) with nonlinear kernels which can properly maintain and utilize the structure of data has been addressed. Some of major challenges of STL can be categorized as follows.

- **High-dimensional tensors**

One of the underlying challenges in STL comes from high dimensionality of tensor objects. This problem will come to notice while trying to use conventional learning algorithms since their assumption is that the instances are represented as vectors. This problem highlights when we trying to reshape tensor to vector where the number of features is extremely high (He et al., 2014).

- **Complex tensor structure**

  Another fundamental problem in STL originates from complex structure of tensor. Traditional tensor- based kernel approaches focus on unfolding tensor data into matrices, where only the one way relationship within in the tensorial data can be preserved. Although, in many real-world applications, the tensorial data have multi-way structures (He et al., 2014).

- **Nonlinear separability**

  According to real-world applications, the data in the input space is usually not linearly divisible. Based on traditional supervised tensor learning methods which can preserve tensor structures are often based upon linear models. Thus these linear methods cannot effectively solve nonlinear learning problems on tensor data (Kong and Yu, 2014), (He et al., 2014).

1.2 Thesis Contribution

This thesis investigates the Supervised Tensor Learning and its application. In particular, solutions for the following problems are proposed:

- **Study of different kernel method and implementation of polynomial kernel within DuSK framework**

  With the aim of better suiting a given dataset, different methods and different kernel functions have been studied. We found that non-linear kernels generally have a better performance on HIV data sets.

- **Diagnosis of HIV infection based on the fMRI and DTI neuroimages of the brain**

  We applied the recent supervised tensor learning methods for the classification of HIV disease on two types of datasets, including fMRI tensor data and DTI brain network data using polynomial in DuSK framework.

- **Extracting intrinsic structure**

  By using CP factorization, we could obtain an elicit intrinsic structure of our data. Which indicates how to explicitly extract multi-mode structure from the original tensorial data.

- **Preserving the multi-mode structure of data in the learning process**

  Different form other tensor methods that damage the multi-mode structure of data by vectorizing them, with our proposed dual mapping method we can preserve the multi-mode structure of data during the learning process.

- **Choice of optimum kernel**

Although we observed a boosting in classification by using non-linear kernels, which can represent the benefit of nonlinear over linear kernels, the choice of an optimum kernel totally depends on the origin of the input data set.

More accurate prediction for HIV infection diagnosis based on fMRI and DTI brain images.

1.3 Thesis Organization

This thesis is organized into three parts and eight chapters. Descriptions of each part and each chapter of the thesis are as follows:

Part I, Preliminaries: This part contains an introduction to the research problem, a review of the existing works and review on fundamental definitions.

Chapter 1: This chapter provides the motivations of this work for readers and indicates the research challenges that the work is focused on. Our approach to these problems along with the list of related contributions is included in this chapter, as well.

Chapter 2: Some backgrounds that will be referenced later in the thesis have been discussed in this chapter. Including methods of tensor decomposition and tensor analysis.

Part II, Models: This part contains problem statement and methodology.

Chapter 3: In this chapter we introduced supervised tensor learning approach and its framework. Some of the most popular STL algorithms also has been discussed in this chapter.

Chapter 4: Main problem of STL has been discussed and formulated at this chapter.

Chapter 5: The logic behind our proposed method has been presented in this chapter along with the related mathematical formulation.

Chapter 6: To compare our method with other similar method we designed several experiments. We include the detail of those experiments along with the observed results in this chapter.

Part III, Conclusion and Future works: This part introduces several approaches that can be used to implement in future based on the proposed method in this thesis.

Chapter 7: A concluding summary of the thesis based on the observed result of the experiments that represented in evaluation section, is presented in this chapter.

Chapter 8: This chapter discusses a few possible extensions of our work and also reviews related open problems. Specifically, the possibility of expanding our method to be able to handle classification and decomposition has been proposed.

1.4 Reader's Guide

The presentation of this thesis is in such a way that each chapter provides preliminary requirement for the next chapters. Considering this fact, it is recommended to be read in the prepared order. However, In some parts of the thesis, a concise preview of the results has been displayed to inspire the approach and offer the reader with additional insight.

Part I
Preliminaries

CHAPTER 2

BACKGROUND AND NOTATION

We represent background references that has been cited and basic definitions that has used in this document in this chapter.

2.1 Notation and Basic Definitions

In this section we list the variables that we will keep referring to in the following chapters:
The conventional notation and definitions in the areas of multilinear algebra, pattern recognition and signal processing (Lu et al., 2008), (Kolda and Bader, 2009), (Savicky and Vomlel, 2007), (De Lathauwer, 1997) has been pursued in this study for readers convince. Thus, in this thesis vectors are denoted by boldface lowercase letters, e.g., **a** matrices by boldface capital letters, e.g., **A**, tensors by calligraphic letters, e.g., $\mathscr{A}$. Their elements are denoted by indices, which typically range from 1 to the capital letter of the index, e.g., $n = 1, ..., N$. To make it more clear, Table I lists the fundamental symbols defined in this study(Hao et al., 2013).

Definition 1(Tensor) : A tensor is a multidimensional array. More specifically, an N-way or Nth-order tensor is an element of the tensor product of N vector spaces, each having its own coordinate system. A third-order tensor has three indices, as shown in Figure 1 (reproduced from (Kolda and Bader, 2009)).

A first-order tensor is called a vector, a second-order tensor is called a matrix, and tensors of a third or higher order are called higher-order tensors. denoted as $\mathscr{A} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ where N is represent the

order of tensor $\mathscr{A}$. The element of $\mathscr{A}$ is shown by $a_{i_1,i_2,\ldots,i_N}$ where i_n is $1 \leq i_n \leq I_n$ and n defines by $1 \leq n \leq N$. Such tensors are the subject of significant research in contemporary computer science (Kolda and Bader, 2009).

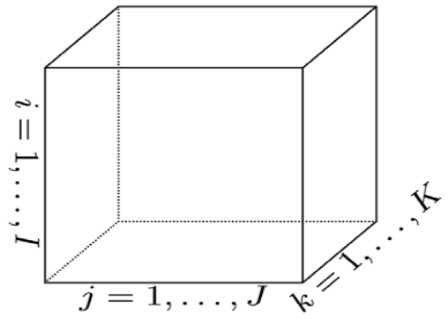


Figure 1: A third order tensor: $\mathbf{x} \in \mathrm{R}^{I \times J \times K}$

Definition 2 (Tensor Product or Outer Product) : The tensor product of a tensor $\mathscr{X} \circ \mathscr{Y}$ of a tensor $\mathscr{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ and another tensor $\mathscr{Y} \in R^{I'_1 \times I'_2 \times \cdots \times I'_M}$ is defined by (Hao et al., 2013).

$$(\mathscr{X} \circ \mathscr{Y})_{i_1,i_2,\cdots,i_N,i'_1,i'_2,\cdots,i'_N} = x_{i_1,i_2,\cdots,i_N} y_{i'_1,i'_2,\cdots,i'_M} \tag{2.1}$$

Definition 3 (Inner Product):The inner product of two same- sized tensors $\mathscr{X}, \mathscr{Y} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ is indicated by the sum of their entries, i.e.,(Hao et al., 2013).

$$\langle \mathscr{X}, \mathscr{Y} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} x_{i_1,i_2,\cdots,i_N} y_{i_1,i_2,\cdots,i_N} \tag{2.2}$$

Definition 4 (n-mode product):The n- mode product of a tensor $\mathscr{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ and a matrix $U \in R^{J_n \times I_n}$ is displayed by $\mathscr{X} \times_n U$ and is a tensor in $R^{I_1 \times I_2 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times I_N}$. Element-wise we have (Hao et al., 2013).

$$(\mathscr{X} \times_n U)_{i_1,i_2,\cdots,i_{n-1},j_n,i_{n+1},\cdots,i_N} = \sum_{i_n=1}^{I_n} x_{i_1,i_2,\cdots,i_N} y_{j_n,i_n} \tag{2.3}$$

Each mode-n fiber is multiplied by the matrix U. N- mode product can also be represented in terms of unfolded tensors:

$$\mathscr{Y} = \mathscr{X} \times_n U \Leftrightarrow Y_{(n)} = U X_{(n)} \tag{2.4}$$

In case when a tensor indicated a multilinear operator, the n-mode product of a tensor with matrix is related to a change of biases. Remark: Given a tensor $\mathscr{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ and a sequence of matrices

$U^{(n)} \in R^{J_n \times I_n}, J_n < I_n, n = 1, \cdots, N$. Projection of tensor $\mathscr{X}$ on the tensor subspace$R^{J_1 \times J_2 \times \cdots \times J_N}$ can be defined as

$$\mathscr{X} \times_1 U^1 \times_2 U^2 \times \cdots \times_N U^N \tag{2.5}$$

Given a tensor $\mathscr{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$, and two matrices $F \in R^{J_n \times I_n}$, $G \in R^{J_m \times I_m}$ one has

$$(\mathscr{X} \times_n F) \times_m G = (\mathscr{X} \times_m G) \times_n F \tag{2.6}$$

Definition 5 (Frobenius Norm):The Frobenius norm of a tensor $\mathscr{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ is the square root of the sum of the squares of all it's elements, i.e., (Hao et al., 2013).

$$\|\mathscr{X}\|_F = \sqrt{\langle \mathscr{X}, \mathscr{X} \rangle} = \sqrt{\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} x_{i_1, i_2, \cdots, i_N}^2} \tag{2.7}$$

Remark: Given two same-sized tensors $\mathscr{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ and, $\mathscr{Y} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ the distance between tensors $\mathscr{X}$ and $\mathscr{X}$ is defined as $\|\mathscr{X} - \mathscr{Y}\|_F$. Note that the Frobenius norm of the difference between two tensors equals to the Euclidean distance of their vectorized representations.(Hao et al., 2013)

Definition 6 (Tensor Rank): Considering tensor $\mathscr{X}$, rank of the tensor $\mathscr{X}$ is defined as the smallest number of rank-one tensors that can produce $\mathscr{X}$ as their sum. Particularly, the smallest number of components in an exact CP decomposition, where exact means that there is equality in the rank of a tensor considered to be rank of tensor $\mathscr{X}$, denoted rank($\mathscr{X}$). Rank decomposition is refer to an ex-

act CP decomposition with $R = rank(\mathscr{X})$components. Although the definition of a tensor and matrix rank is the same the properties of it is competently different. The definition of tensor rank is an exact analogue to the definition of matrix rank, but the properties of matrix and tensor ranks are quite different.

M	The total number of tensor samples
L	The number of classes
$\{X_m, y_m\}_{m=1}^{M}$	A set of tensor samples
X_m	The mth input tensor sample
y_m	The label of Xm
N	The order of $\mathscr{X}_m \in R^{I_1 \times I_2 \times \cdots \times I_N}$
$R = rank\,(X_m)$	The rank of X_m
W	The weight parameter
b	The bias variable
C	The trade-off parameter
ξ	The slack variables
α, β	The Lagrange multipliers
ε	The threshold parameter
$w^{(1)} \circ w^{(2)} \circ \cdots w^{(N)}$	Rank-1 tensor of nmode product of XW
$X \times_n w$	and w
$\|.\|_F$	Frobenius norm

2.2 A Brief Review of Tensor Decomposition Techniques

This section provides a brief review on higher-order tensors decomposition algorithms and mathematics. On the next section a general overview of higher-order tensors analysis methods for classification is going to be introduced (Kolda and Bader, 2009).

Decomposition also known as Matricization, unfolding or flattening, is the procedure of rearranging the elements of an N-way array into matrix or vector. Two of the most popular decomposition methods are Tucker and CP decomposition. During this project we used CP decomposition method to decompose our tensorial data. Int he future works chapter we also proposed the replacement of Tucker decomposition with CP decomposition in our proposed method based on the input data set. Therefore we are going to have a brief review on these two techniques.

- **CP Decomposition**

  Psychometrics can be considered where CP decomposition originated in 1970. Carroll and Chang (Carroll and Chang, 1970) introduced CANDECOMP in the concept of multiple analysis similarity or dissimilarity matrices from a various of subjects. The idea was averaging over the data for all the subjects can obliterate different points of view. Harshman (Harshman, 1970)proposed PARAFAC because it eradicates the vagueness associated with two-dimensional Principle Component Analysis (PCA) and thus has better uniqueness properties.

  As mentioned earlier in definition of tensor rank, there is no finite algorithm for determining the rank of a tensor (Håstad, 1990). Thus, the first challenge that arises in calculating a CP

decomposition is how we can choose the number of rank-one elements. In most cases, multiple CP decomposition with different numbers of components fits until one is good. In ideal situation, where the data are noise-free, we have a procedure for computing CP with a given number of elements, then we can do that calculation for R = 1,2,3,... elements and set the stopping point where we get the first value of R that gives us a fit of 100%. Although, there might be many complication in this method but we need to consider that there is still no perfect procedure to find a CP for a given number of components.

Considering all the above explanation we can define CP decomposition in general as below: When tensor $\mathscr{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is given and an integer Rdenotes the rank, if $\mathscr{A}$it can be expressed as

$$\mathscr{A} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \otimes \mathbf{a}_r^{(2)} \otimes \cdots \otimes \mathbf{a}_r^{(N)}, \tag{2.8}$$

we call it CP factorization (see Figure 2 reproduced from (Kolda and Bader, 2009)for graphical representations). To be convenient, in the following we write $\prod_{n=1}^{N} \otimes \mathbf{a}^{(n)}$ for $\mathbf{a}^{(1)} \otimes \mathbf{a}^{(2)} \otimes \cdots \otimes \mathbf{a}^{(N)}$.

Many researchers have previously used CP decomposition in neurimaging. The author of (Mocks, 1988) independently discovered CP in the area of event-related potentials in brain imaging; his work was the beginning of using CP in brain imaging. He also included results on uniqueness, remarks and how to find the number of factors. Then, Andersen and Rayens (Andersen and Rayens, 2004) applied CP decomposition to fMRI data set arranged as voxels by time by run and

also as voxels by time by trial by run. Their work was inspiring for many researchers who later worked on fMRI brain imaging to use CP decomposition.

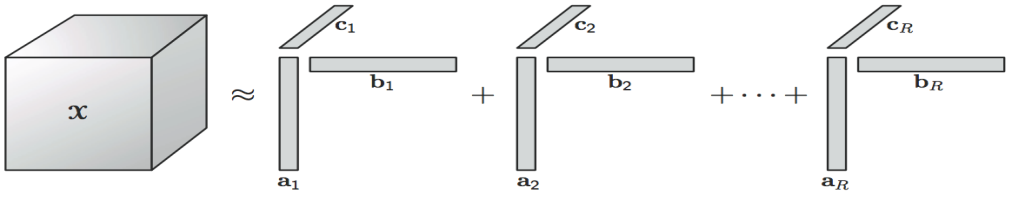


Figure 2: CP decomposition of a three-way array

- **Tucker Decomposition** Tucker decomposition is defined (see Figure 3 reproduced from(Kolda and Bader, 2009)) as a form of higher-order PCA. It factorizes a tensor into a multiplication of a core tensor by a matrix along each mode as follows:

$$\mathscr{X} \approx \mathscr{C} \times_1 U_1 \cdots \times_N U_N, \tag{2.9}$$

$$= \sum_{r_1=1}^{R_1}, \cdots, \sum_{r_N=1}^{R_N} c_{r_1 \cdots r_N} \mathbf{u}_{r_1} \circ \cdots \circ \mathbf{u}_{r_N} \tag{2.10}$$

$$\triangleq \|\mathscr{C}; \mathbf{U_1}, \mathbf{U_2}, \cdots, \mathbf{U_N}\|$$

where $\mathscr{X} \in R^{I_1 \times I_2 \times \cdots \times I_N}$ of a tensor is the original tensor. $\mathbf{U}_1 \in \mathbb{R}^{I_1 \times R_1}, \cdots, \mathbf{U}_N \in \mathbb{R}^{I_N \times R_N}$ are the factor matrices, where usually considered columnwise orthogonal, and can be represented as the principle components in each mode. $\mathscr{C} \in R^{R_1 \times R_2 \times \cdots \times R_N}$ is called the core tensor, which stands for all probable linear interactions between the components of each mode.(Wu et al., 2013)

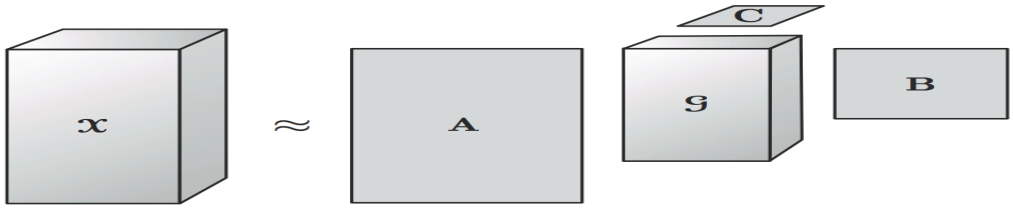


Figure 3: Tucker decomposition of a three-way array

2.3 A General Overview of Tensor Analysis

Tensor representation is useful for minimizing the small sample size (SSS) problem in discriminative subspace selection. This usefulness is rooted in the fact that the structure information of objects in computer vision research provides a means of reducing the number of latent parameters used to represent a learning model. Therefore, we apply this structure information to vector-based learning and then generalize that learning to tensor-based learning as a supervised tensor learning (STL) framework, which accepts tensors as input.

To obtain the STL solution, an alternating projection optimization procedure is applied. The STL framework is a combination of convex optimization and the operations in multilinear algebra. Tensor modeling is useful in reducing the overfitting problem in vector-based learning. Based on STL and its alternating projection optimization procedure, support vector machines (SVM), the minimax probability machine, Fisher discriminant analysis, and distance metric learning are generalized to respectively support tensor machines, the tensor minimax probability machine, tensor Fisher discriminant analysis, and multiple distance metrics learning. The iterative procedure for feature extraction within STL is also of interest. To evaluate the effectiveness of STL, the tensor minimax probability machine can be implemented for image classification. Through comparison with the minimax probability machine, it can be concluded that the tensor version reduces the overfitting problem (Hao et al., 2013).

Within the past 20 years, researchers have given strong attention to data representation methods such as tensor decomposition (Shashua and Levin, 2001) (Bourennane et al., 2010) and multilinear subspace learning in order to address the curse of dimentionality and SSS problems (Lu et al., 2008), (Geng et al., 2011).

Currently, in order to generalize the SVM learning framework to tensor patterns, researchers (Tao et al., 2005) have proposed developing multilinear models (Hao et al., 2013).

For example, a supervised tensor learning (STL) framework was represented by Tao et al. (Tao et al., 2005). They applied their method as a combination of convex optimization and multilinear operators in which the weight parameter was unfolded into a vectors (Shashua and Levin, 2001). Using this learning framework, Cai et al. (Cai et al., 2006) studied second-order tensors and presented a linear tensor least square (TLS) classifier. Tao et al. (Tao et al., 2005) extended the classical linear C-support vector machine (C-SVM), (V-SVM), and least squares SVM to general tensor patterns. Based on the SVM technique within the STL approach, Liu et al. (Liu et al., 2008) applied dimentionality reduction to tensors as input for video analysis. In addition, Kotsia et al. (Kotsia and Patras, 2011) generalized Tucker decomposition of the weight parameter instead of the vectors to preserve more structural information. Furthermore, instead of using the classical maximum-margin criterion, Wolf et al used the orthogonality constraints on the columns of the weight parameter. (Wolf et al., 2007) attempted to minimize the rank of the weight parameter, and Pirsiavash et al. (Pirsiavash et al., 2009) loosened the orthogonality constraints to further enhance the Wolfs method (Hao et al., 2013).

at this section we are going to review classification methods starting with SVM and then we will have a brief review on other similar classification method that take tensor as an impute different form other algorithms.

In (Hao et al., 2013) Hao, He, Chen, and Yang comprehensively reviewed studies of support vector machines for extension to tensors. Their review formed the basis for much of the information that follows.

Currently, most learning machines, such as support tensor machines (STM), contain nonconvex optimization and need to employ iterative techniques. Clearly, use of iterative algorithms has two shortcomings: it is time-consuming and is negatively affected by local minima. An innovative linear support higher-order tensor machine (SHTM) has been represented in as overcoming iterative algorithms defects (Hao et al., 2013).

From a theoretical perspective, SHTM extends the linear C-SVM to tensor patterns. When vectors are the input patterns, SHTM degenerates into the standard C-SVM. To illustrate SHTM performance, experiments have been performed on nine second-order face recognition datasets and three third-order gait recognition datasets. In these experiments, a statistic test showed that in comparison with STM and C-SVM with the RBF kernel, SHTM provided significant performance gain with respect to test accuracy and training speed, especially for higher-order tensors (Hao et al., 2013).

In past decades, numerous advanced algorithms have been suggested and have been very successful in multiple applications. Among these, the most promising example is the SVM (Vapnik, 2013), which is especially useful for pattern recognition, computer vision, and image processing due to various theoretical and computational advantages.

However, the standard SVM model cannot directly use non-vector patterns because it is based on vector space, while real-world image and video data are more appropriately represented as matrices (second- order tensors) or higher-order tensors. For instance, color video sequences are typically displayed as fourth-order tensors. Tensor objects can be reformulated into vectors to meet SVM input requirements. However, previous studies have shown that this direct reformulation disrupts the genuine

structure and correlation in the original data and results in the curse of dimensionality and small sample size SSS problems and thus poor SVM performance is observed (Li et al., 2006) (Hao et al., 2013).

Part II
Models

CHAPTER 3

SUPERVISED TENSOR LEARNING

3.1 A Brief Introduction to Supervised Tensor Learning (STL)

Tensorial representation of data is useful to decrease SSS problem effect in discriminative subspace selection. According to (Tao et al., 2005) this is basically because the structure information of objects in computer vision research can be used as a logical constraint to reduce the number of unknown features used to represent a learning model. Hence, this information applied to the vector-based learning and generalize the vector-based learning to the tensor-based learning as the supervised tensor learning (STL) algorithm, which accepts tensors as input(Tao et al., 2005).

The alternating projection optimization framework is expanded, in order to reach the solution of STL. The STL framework can be defined as a mixture of the operations in multilinear algebra and the convex optimization. The tensor representation is helpful to decrease the overfitting problem in vector-based learning. We generalized all vector based models like SVM, minimax probability machine, Fisher discriminant analysis, and distance metric learning, to tensors like STM, tensor minimax probability machine, tensor Fisher discriminant analysis, and the multiple distance metrics learning, respectively, based on STl optimization procedure. The iterative procedure for feature extraction within STL has also been reviewed in this study. The tensor minimax probability machine for image classification has been implemented to evaluate the effectiveness of STL and as result the tensor framework decreased the overfitting problem in compare with minimax probability machine(Tao et al., 2005).

As He et al. (2014) point out in(He et al., 2014) , supervised learning can be considered one of the bases of fundamental data mining tasks. In supervised learning, the typical approaches generally assume that data instances are demonstrated as feature vectors. In many real-world applications, however, data instances are more naturally displayed as second-order (matrices) or higher-order tensors, where the number of modes or ways corresponds to the tensor order. Supervised learning with this type of data is referred to as STL, in which each instance in the input space is displayed as a tensor. Given the great increase in tensorial data, STL has recently drawn significant attention in in the machine learning and data mining areas(He et al., 2014).

Among many advances in data analysis technology, tensor data is becoming increasingly important in many applications. However, the problem of STL is a critical issue in the data mining and machine learning fields. Conventional approaches for STL generally concentrate on learning kernels by flattening the tensor into vectors or matrices, with the disadvantage of losing the structural information within the tensors. He et al. (2014) introduced a new method for designing structure-preserving kernels for STL. To encode prior knowledge in a kernel, they proposed using the natural structure within the tensorial representation (He et al., 2014).

3.2 STL Popular Algorithms

In this section three of the most popular algorithms in STL approach, where two of them has been first introduced in background chapter, are going to be discussed.

The supervised learning algorithms that are going to be discussed here, different form previous data classification method which accepts only vectors as an input, can directly accept tensors as an input.

- **Least Square (TLS)**

  Least Square classifier is one of the most common classifiers. In order to develop the TLS model, th idea of least square has been expanded tensor space in (Cai et al., 2006). The objective function of TLS defines as follows:

$$\min_{\mathbf{u},\mathbf{v}} \sum_{i=1}^{m} \left(\mathbf{u}^T \mathbf{X}_i \mathbf{v}_- y_i\right)^2 \tag{3.1}$$

- **Support Tensor Machine (STM)**

  SVM comes from the category of pattern classification algorithms developed by (Vapnik, 2013). STM is basically the extension of SVM to tensor space. Detailed mathematical formulation and objective function has been presented in problem statement chapter.

- **Support Higher- order Tensor Machine (SHTM)**

  SHTM model introduced in (Hao et al., 2013). First, the STM model using concepts of multilinear algebra has been reformulated and a base model obtained, which happens to be identical to the linear C SVM model, but is an independent implementation from a multilinear algebra viewpoint. Second, rank-one tensor decomposition integrated into the base model and present SHTM

optimization model. More detail on formulation can be found in (Hao et al., 2013). Since this method is also using a Dual Structure-preserving mapping with linear kernel, another name of this method is $DuSK_{Linear}$. We choose this method as one of our baselines to compare our method with it.

CHAPTER 4

PROBLEM STATEMENT

In this chapter we are going to explain the STL optimization problem with detail and at the next chapter we are going to present the method that we suggested to solve this problem. This chapter and the next two chapter are mainly designed based on th eproposed method by (He et al., 2014) and their experiments.

4.1 STL Optimization Problem Statement

Tensorial representation of data is useful to decrease SSS problem effect in discriminative subspace selection. According to (Tao et al., 2005) this is basically because the structure information of objects in computer vision research can be used as a logical constraint to reduce the number of unknown features used to represent a learning model. Hence, this information applied to the vector-based learning and generalize the vector-based learning to the tensor-based learning as the supervised tensor learning (STL) algorithm, which accepts tensors as input (Tao et al., 2005).

The alternating projection optimization framework is expanded, in order to reach the solution of STL shortcomings. The STL framework can be defined as a mixture of the operations in multilinear algebra and the convex optimization. The tensor representation is helpful to decrease the overfitting problem in vector-based learning. SVM, minimax probability machine, Fisher discriminant analysis, and distance metric learning methods has been generalized based on, STL and its alternating projection optimization

procedure to STM, tensor minimax probability machine, tensor Fisher discriminant analysis, and the multiple distance metrics learning, respectively.The iterative procedure for feature extraction within STL has also been reviewed in this study (Tao et al., 2005).

As pointed out (He et al., 2014), supervised learning can be considered one of the bases of fundamental data mining tasks. In supervised learning, the typical approaches generally assume that data instances are demonstrated as feature vectors. In many real-world applications, however, data instances are more naturally displayed as second-order (matrices) or higher-order tensors, where the number of modes or ways corresponds to the tensor order. Supervised learning with this type of data is referred to as STL, in which each instance in the input space is displayed as a tensor. Given the great increase in tensorial data, STL has recently drawn significant attention in in the machine learning and data mining areas (He et al., 2014).

Among many advances in data analysis technology, tensor data is becoming increasingly important in many applications. However, the problem of STL is a critical issue in the data mining and machine learning fields. Conventional approaches for STL generally concentrate on learning kernels by flattening the tensor into vectors or matrices, with the disadvantage of losing the structural information within the tensors. He et al. (2014) introduced a new method for designing structure-preserving kernels for STL. To encode prior knowledge in a kernel, they proposed using the natural structure within the tensorial representation which is also the method that has been used and expanded at this study (He et al., 2014).

Assuming a training set that contains M pairs of samples $\{\mathscr{X}_i, y_i\}_{i=1}^{M}$ for binary tensor classification problem, where $\mathscr{X}_i \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ are the input of the sample and coressponding class lables of $\mathscr{X}_i$are $y_i \in \{-1, +1\}$.(Hao et al., 2013), was denoted that the problem of tensor classification can be expressed

as a quadratic convex optimization problem in the linear SVM framework. Based on the result represented in (Hao et al., 2013), it has been shown in this document that how it can be modeled as a kernel learning problem. Suppose we are given the optimization problem of linear tensor classification as"

$$\min_{\mathscr{W},b,\xi} \frac{1}{2}\|\mathscr{W}\|_F^2 + C\sum_{i=1}^{M}\xi_i, \tag{4.1}$$

$$\text{s.t. } y_i\left(\langle \mathscr{W}, \mathscr{X}_i\rangle + b\right) \geq 1-\xi_i, \tag{4.2}$$

$$\xi_i \geq 0, \forall i = 1,\cdots,M. \tag{4.3}$$

Where the weight tensor of the separating hyperplane is $\mathscr{W}$, b is the bias, ξ_i is the error of the ith training sample, and C is the trade-off between the classification margin and miss-classification error.(He et al., 2014)

The generalization of the standard linear SVM problem to tensor patterns in tensor space, is presented in optimization problem in (4.1)-(4.3). When the input samples $\mathscr{X}_i$ are vectors, it becomes the standard linear SVM. By introducing a nonlinear feature mapping $\phi : \mathbf{x} \to \phi(\mathbf{x}) \in \mathfrak{H} \subset \mathbb{R}^H$, we develop a nonlinear extension of (4.1)-(4.3) based on the kernel method for the expansion of linear SVM to the nonlinear SVM. Although it is critical to drive the same model for tensor-based kernel learning.

Assume a tensor $\mathscr{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$, which is mapped into the Hilbert space $\mathfrak{H}$ by

$$\phi : \mathscr{X} \to \phi(\mathscr{X}) \in \mathbb{R}^{H_1\times H_2\times\cdots\times H_P}. \tag{4.4}$$

It is interesting to know that, the project tensor $\phi(\mathscr{X})$ in space $\mathfrak{H}$ might have different order with $\mathscr{X}$, and depending on the feature mapping function$\phi(.)$ each mode dimension is higher even an infinite dimension may happen. . Such a Hilbert space, is called the high-dimensional tensor feature space or merely a tensor feature space.The following model in this space, generated based on the same principle as the generation of linear classification model in the original tensor space:

$$\min_{\mathscr{W},b,\xi} \frac{1}{2}\|\mathscr{W}\|_F^2 + C\sum_{i=1}^{M}\xi_i, \tag{4.5}$$

$$\text{s.t. } y_i\left(\langle \mathscr{W}, \phi(\mathscr{X}_i)\rangle + b\right) \geq 1-\xi_i, \tag{4.6}$$

$$\xi_i \geq 0, \forall i = 1,\cdots,M. \tag{4.7}$$

(4.5)-(4.7)considered to be a linear model from the point of view of high- dimensional tensor. But, from the viewpoint of the original tensor space, it considers as a nonlinear model. As mentioned above, when the input samples $\mathscr{X}_i$ are vectors, the model turns into the standard nonlinear SVM. If the feature mapping function $\phi(.)$ is an identical function, i.e., $\phi(\mathscr{X}) = \mathscr{X}$, it can be the same as (4.1)-(4.3) . Therefore, we say that the optimization model (4.5)-(4.7) is the nonlinear equivalent of (4.1)-(4.3).

Through (4.8)-(4.10) we demonstrated how this model can be exploited to achieve tensor-based kernel optimization model. Using Lagrangian relaxation method presented in (Chong and Zak, 2013), it is easy to check that the dual problem of (4.5)-(4.7) is

$$\max_{\alpha_1,\alpha_2,\cdots,\alpha_M} \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j \langle \phi(\mathscr{X}_i), \phi(\mathscr{X}_j) \rangle \tag{4.8}$$

$$\text{s.t.} \sum_{i=1}^{M} \alpha_i y_i = 0, \tag{4.9}$$

$$0 \leq \alpha_i \leq C, \forall i = 1, \cdots, M. \tag{4.10}$$

Where α_i are the Lagrangian multipliers and $\langle \phi(\mathscr{X}_i), \phi(\mathscr{X}_j) \rangle$ are the inner product between the mapped tensors of $\mathscr{X}_i$ and $\mathscr{X}_j$ in the tensor feature space.

During (4.8)- (4.10) training data only appear in the form of inner products which is the advantage of formulation of (4.8)- (4.10) over (4.5)-(4.7). Based on the fundamental principle of kernel method, by substituting the inner product $\langle \phi(\mathscr{X}_i), \phi(\mathscr{X}_j) \rangle$ with a suitable tensor kernel function $\kappa(\mathscr{X}_i, \mathscr{X}_j)$, we will get the tensor-based kernel model. The resulting decision function presented in (4.11)

$$f(\mathscr{X}) = sign\left(\sum_{i=1}^{M} \alpha_i y_i \kappa(\mathscr{X}_i, \mathscr{X}) + b \right). \tag{4.11}$$

CHAPTER 5

METHODOLOGY

In this chapter, different types of DuSK for STL with applications to neuroimages are introduced. This chapter focuses on DuSK mathematical formulation, and in the evaluation chapter we are going to concentrates on comparing DuSK performance with several baselines.

As mentioned in previous chapters, structural information will be lost during flattening of tensorial data to the matric or vectors to be used as an input for conventional kernels. Flattening of tensorial data makes conventional methods prone to overfitting, specifically in case of the SSS problem. DuSK offers a tensor kernel-based method that can protect the tensorial structure of data through dual tensor mapping. Using the dual tensor mapping function, while preserving tensorial structure of data, DuSK is mapping each tensor in the input space to another tensor in the feature space. From a theoretical point of view, the DuSK method can be considered as a generalization of the traditional kernels in the vector space to tensor space.

5.0.1 A Dual Structure-preserving Kernel (DuSK)

From the decision function proposed in the problem statement chapter, we see that tensor-based kernel learning becomes a study of kernel function, and the effectiveness of kernel methods completely depends on the data representation encoded into the kernel function. The aim of the method proposed in this thesis is to preserve the natural structure of the tensor in order to achieve better kernel learning (He et al., 2014).

Tensors are employed to provide a genuine and sufficient representation for multiway data, but such a data representation is not necessarily appropriate for kernel learning. Like the earlier analysis for tensor object characteristics, the vital information in a tensor incorporated in to its multiway structure. Therefore, an important feature of kernel learning for complex objects is to display them by key chains of structural characteristics that are simpler to manipulate, and to structure kernels on such sets.

Based on the mathematical definition of a tensor, "we can better understand the structure of the tensor when that tensor object has the reflex of product structure."Previous research revealed that CP factorization is effective for eliciting this structure.Given these findings, we consider how to take advantage of the benefits of CP factorization to train a structure preserving kernel in tensor product feature space.

In so doing, we display each each tensor object as a summation of rank-one tensors in the original space and we then map them into the tensor product feature space to achieve our kernel learning. Bellow we explained how we designed the feature mapping.

We start by defining the following mapping on a rank-one tensor $\prod_{n=1}^{N} \otimes \mathbf{x}^{(n)} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$.

$$\phi : \prod_{n=1}^{N} \otimes \mathbf{x}^{(n)} \rightarrow \prod_{n=1}^{N} \otimes \phi \left(\mathbf{x}^{(n)} \right) \in \mathbb{R}^{H_1 \times H_2 \times \cdots \times H_N}. \tag{5.1}$$

Let the CP factorization of $\mathscr{X}, \mathscr{Y} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ be $\mathscr{X} = \sum_{r=1}^{R} \prod_{n=1}^{N} \otimes \mathbf{x}_r^{(n)}$ and $\mathscr{Y} = \sum_{r=1}^{R} \prod_{n=1}^{N} \otimes \mathbf{y}_r^{(n)}$ respectively. Based on the kernel function concept the kernel function the kernel can be directly defined with inner product in the feature space. Therefore, when $R = 1$, based on the mapping presented above and Eq. **??**, the naive tensor product kernels can be directly derived as follows:

$$\kappa(\mathscr{X},\mathscr{Y}) = \prod_{n=1}^{N} \kappa\left(\mathbf{x}^{(n)},\mathbf{y}^{(n)}\right). \tag{5.2}$$

Nevertheless, many researchers have shown that a simple rank-one tensor cannot offer a compressed and informative diplay of original data (Yada Zhu and Lawrence, 2012). The question remains of how to develop feature mapping when R is more than one.

According to the definition of the kernel function, the feature space consist of a high-dimensional space of the original space, that follows the same operation rules. Therefore, tensor data can be directly factorized in the feature space in the same way as in the original space. This approach is equivalent to the following mapping:

$$\phi : \sum_{r=1}^{R}\prod_{n=1}^{N}\otimes\mathbf{x}_r^{(n)} \rightarrow \sum_{r=1}^{R}\prod_{n=1}^{N}\otimes\phi\left(\mathbf{x}_r^{(n)}\right). \tag{5.3}$$

This approach corresponds to mapping tensors into high-dimensional tensor space that retaining the original structure. To be more specific, this approach can be viewed as as mapping the original data into a tensor feature space and performing CP factorization in in that space. This mapping process is refer to as the dual-tensorial mapping function (see Figure 4 modified from (He et al., 2014)).

Once, the CP factorization of the data is mapped into the tensor product feature space, the kernel is simply defined as the standard inner product of tensors in that feature space. The result is as follows (He et al., 2014):

$$\begin{aligned}\kappa\left(\sum_{r=1}^{R}\prod_{n=1}^{N}\otimes\mathbf{x}_r^{(n)}\ ,\ \sum_{r=1}^{R}\prod_{n=1}^{N}\otimes\mathbf{y}_r^{(n)}\right)\\ =\sum_{i=1}^{R}\sum_{j=1}^{R}\prod_{n=1}^{N}\kappa\left(\mathbf{x}_i^{(n)},\mathbf{y}_j^{(n)}\right)\end{aligned} \tag{5.4}$$

Based on the derivation procedure, a kernel of this type can preserve the multi-way structure's flexibility in calculations. In general, (He et al., 2014) extends traditional kernels in the vector space to tensor space. Consequently each vector kernel can be employed in this framework for STL in conjunction with kernel machines (He et al., 2014).

5.0.2 DuSK with different Kernels

Next we address the case in which we used the polynomial kernel in our framework. This type of kernel is widely used and has been shown to be effective in a various of contexts. Let us assume that a set of tensor data is $\{(\mathscr{X}_i,\ y_i)\}_{i=1}^{M}$, where $\mathscr{X}_i \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$is the case. When we want to use different kernel we need to reformulate the Eq. 5.4. In continue we present our proposed DuSK with polynomial kernel and already used DuSK with RBF and Linear kernel.

- **DuSK with Polynomial Kernel**

  By replacing polynomial kernel in Eq. 5.4 we get Eq. 5.5

$$\kappa(\mathscr{X},\mathscr{Y})=\sum_{i=1}^{R}\sum_{j=1}^{R}\prod_{n=1}^{N}\kappa\left(\mathbf{x}_i^{(n)^T}.\mathbf{y}_j^{(n)}+c\right)^d \tag{5.5}$$

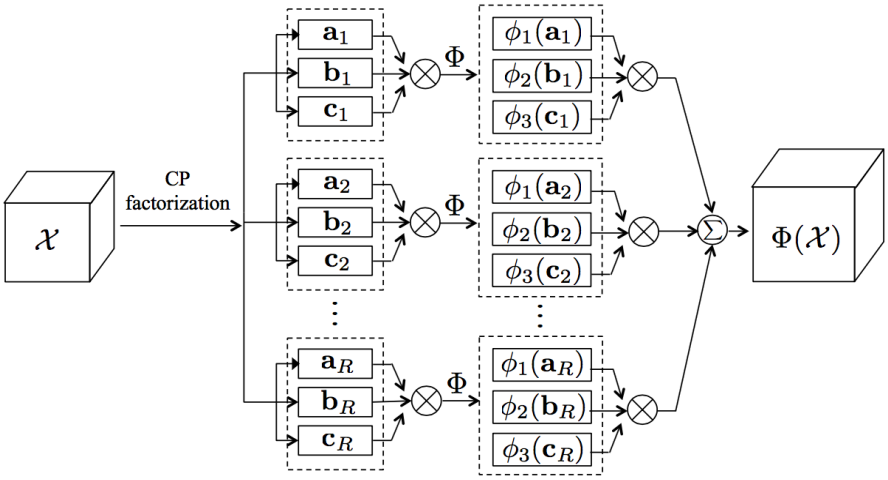


Figure 4: Dual-tensorial mapping procedure

In Eq. 5.5 d denoted the degree of polynomial kernel where d is d= 2 we have quadratic polynomial. Quadratic polynomial is very popular since using higher d might result in overfitting specially in case of natural language problems (NPL). Free parameter $c \geq 0$ is a trading off that reflect the effect of higher-order versus lower-order terms in the polynomial. When c = 0, the kernel is called homogeneous.

- **DuSK with RBF kernel**

  In this case by constituting Gaussian RBF kernels, we can rewrite the Eq. 5.4 as Eq. 5.6

$$\begin{aligned}\kappa(\mathscr{X},\mathscr{Y}) &= \sum_{i=1}^{R}\sum_{j=1}^{R}\prod_{n=1}^{N}\kappa\left(\mathbf{x}_i^{(n)},\mathbf{y}_j^{(n)}+c\right) \\ &= \sum_{i=1}^{R}\sum_{j=1}^{R}\exp\left(-\sigma\sum_{n=1}^{N}\|\mathbf{x}_i^{(n)}-\mathbf{y}_j^{(n)}\|^2\right)\end{aligned} \tag{5.6}$$

where σ is used to set an proper bandwidth. We denote this kernel as $DuSK_{RBF}$ which first proposed in this (He et al., 2014) paper.

- **DuSK with Linear kernel**

  DuSK with Linear kernel has been introduced in (Hao et al., 2013) paper for the first time. In this method they replaced the kernel function in Eq. 5.4 by linear kernel you can see the the main equation for deriving $DuSK_{Linear}$ below in Eq. 5.7:

$$\kappa(\mathscr{X},\mathscr{Y}) = \sum_{i=1}^{R}\sum_{j=1}^{R}\prod_{n=1}^{N}\kappa\left\langle\mathbf{x}_i^{(n)}.\mathbf{y}_j^{(n)}\right\rangle \tag{5.7}$$

CHAPTER 6

EVALUATION

6.1 Data Collection and Preprocessing

This chapter evaluates the performance of the proposed models on HIV fMRI tensor and DTI matric dataset.

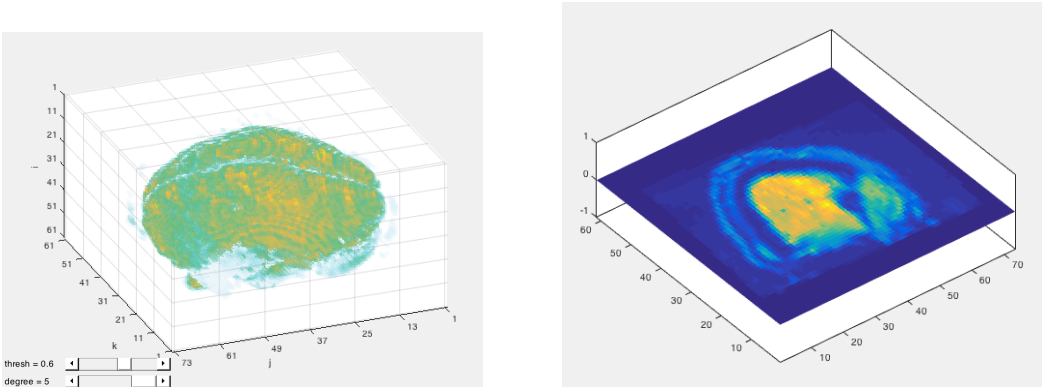
The study here reported seeks to critically analyze supervised tensor learning methods designed for the medical diagnosis of HIV infection. We use one real-world functional Magnetic Resonance Imaging (fMRI) tensor dataset and one real-world Diffusion Tensor Imaging (DTI) matrix dataset in our experimental evaluation as follows.

- **fMRI tensor data:**

  fMRI is a short form for denoting Functional Magnetic Resonance Imaging. This data set is collected from the Department of Radiology in Northwestern University (Wang et al., 2011). It contains fMRI brain imaging data of a total 83 patient where 25 patients with early HIV infection (positive) and 58 normal controls (negative).

  Based on (He et al., 2014) to preprocess the fMRI tensor data we applied SPM8 toolbox. The first ten volumes for each sample has been deleted, and functional images were rearranged to the first volume, slice timing corrected, and normalized to the MNI format and dimentionaly flatted with an 8-mm FWHM Gaussian kernel. Resting-State fMRI Data Analysis Toolkit (REST) was

then used to remove the linear trend of time series and temporally band-pass filtering (0.01 0.08 Hz)(Kong et al., 2013). The average value of each subject over time series was computed within each of those boxes, thereby resulting in 83 samples and a sum total of 61*73*61 = 271633 voxels (or features) (He et al., 2014). Vitalization of HIV fMRI data presented in Figure 5a and Figure 5b.



(a) fMRI Tensorial visualization

(b) A slice of fMRI image

Figure 5: brain fMRI data visualization

- **DTI brain network data:**

DTI is a short form of Difussion Tensor Imaging. This dataset is collected from the Chicago Early HIV Infection Study at Northwestern University (Wang et al., 2011). It involves DTI images of 77

subjects, 56 of which are early HIV patients (positive) and the other 21 subjects are seronegative controls (negative).

To preprocess the DTI images and constructed the brain networks we used FSL toolbox absed on (He et al., 2014). We perform preprocessing using the standard process as demonstrated in Figure 6. The preprocessing corrects distortion, filter noises, sample repetitive from the distributions of principal diffusion directions for each voxel. To elicit the brain networks from DTI images, we focus on the 116 anatomical volumes of interest (AVOI), each of which displayed a special part of the brain, and extract a sequence of reaction from them. We parceled each DTI image into the 90 cerebral regions via the propagation of the Automated Anatomical Labeling (AAL), thereby resulting in 77 samples with the 90*90 feature matrix (graph). Each node in graph represents a brain region, and links are created based on the correlations between different brain regions(Kong et al., 2013). In Figure 6 a visualization of DTI brain network has been showed.

- **Difference between fMRI and DTI brain network images:**

  fMRI Images is using to observe activities of specific parts of the brain while DTI imaging use to find relation between different parts of the brain to fins a network based on the parts that communicating. Their application varies based on the disease and physicians approach. But the main application of these types of medical imaging is in diagnosis and treatment of nervous system related disease.

Figure 6: DTI brain network vitualization

6.2 Baselines and Metrics

- $DuSK_{Linear}$**:** $DuSK_{Linear}$ or Linear SHTM is a linear support higher-order tensor machine (Hao et al., 2013), which is one of the most effective methods for tensor classification that generalizes linear SVM to tensor pattern using CP factorization and can be regarded as a special case of (**?**)e2014dusk, namely the constituent kernels are linear kernels. This baseline along with DuSK-RBF are used to test the ability of our proposed method to match with complex (possibly nonlinear) structured data.

- $DuSK_{RBF}$**:** $DuSK_{RBF}$ is an extension to DuSK algorithm, the constituent kernels is RBF kernel, which is similar to DuSk linear is one of the most effective methods for tensor classification that generalizes linear SVM to a tensor based method using CP factorization (**?**)e2014dusk.This base-

line along with DuSK- linear are used to test the ability of our proposed method to implement on complex (possibly nonlinear) structured data (He et al., 2014).

- **Linear SVM:** Linear SVM has also been increasingly used to handle fMRI data. In some cases, it performs better than SVM using nonlinear kernels (He et al., 2014).

- SVM_{RBF}**:** SVM_{RBF} is a Gaussian-RBF kernel-based SVM, which is now the most widely used vector-based method for classification. In the following methods, if not stated explicitly, we use SVM with Gaussian RBF kernel as the classifier(He et al., 2014).

6.3 Experimental result on HIV Neuroimaging data set

The effectiveness of an algorithm always analysis by testing accuracy.Therefore, accuracy has been used as a metric in this study to evaluate the performance of the proposed algorithm. Results of our experiment has been presented in 3 tables.The accuracy results displayed in table 1 and 2 are the average accuracy of CP decomposition with rank R where R is R= $\{1,2,3,4,5,6,7,8\}$.

- **Classification Performance** We designed and performed our experiments based on (He et al., 2014). We first randomly divided samples of the whole data to 80% as a training, and consider the rest of samples as the test set. We repeated this random sampling experiment 50 times for all methods and reported the average performances of each method. The average classification accuracy and standard deviation of seven algorithms on three data sets displayed in Table I and II. We highlighted the best results in bold type.

- **Result Analysis**

According to the results of experiments displayed in Table I and II, we denoted that the classification accuracy of each method on different data set can be quite different. However, the best method that outperforms other methods especially for DTI data set is our proposed method. It should be notice that in neuroimages area it is very hard to achieve even moderate classification accuracy on these data sets since this data has extremely high dimensionallity but with a amount of samples that can cause SSS problem.

It has been represented that our proposed method was sufficient on DTI data classification. As demonstrated, CP factorization can fully capture the multi-way structure of the data, thus our method consider this fact in the learning process.

Result of the experiments presented in below tables:

TABLE I: AVERAGE CLASSIFICATION ACCURACY COMPARISON: MEAN (STANDARD DEVIATION)

Dataset	$DuSK_{Poly}$	$DuSK_{Linear}$	$DuSK_{RBF}$	SVM_{RBF}	Linear SVM
fMRI Tensor data	$0.69\,(0.03)$	$0.70\,(0.01)$	$\mathbf{0.74}\,(\mathbf{0.00})$	$0.70\,(0.00)$	$0.74\,(0.01)$

Results that showed in the Table I demonstrated that $DuSK_{Poly}$ has almost similar performance to other baselines and does not make a better performance on three dimensional fMRI tensorial data set in compare to presented baselines.

From observation of the experimental results on DTI data set that presented in Table II, we can indicate $DuSK_{Poly}$ performs better than all of the baselines. According to the above table $DuSK_{Poly}$ can boost the classification accuracy of DTI matric images.

TABLE II: AVERAGE CLASSIFICATION ACCURACY COMPARISON: MEAN (STANDARD DEVIATION)

Data set	$DuSK_{Poly}$	$DuSK_{Linear}$	$DuSK_{RBF}$	SVM_{RBF}	Linear SVM
DTI Matric images	$\mathbf{0.710}\,(\mathbf{0.018})$	$0.671\,(0.033)$	$0.535\,(0.027)$	$0.686\,(0.023)$	$0.671\,(0.020)$

According to the presented results in Table Iand Table II we can state that nonlinear kernel learning can prompt the accuracy of our classifier on HIV fMRI and DTI data sets.

Part III
Conclusion

CHAPTER 7

CONCLUSION

This concluding chapter, summarizes the results of the thesis.

In this thesis, we proposed a new tensor-based kernel approach that first operates directly on tensors. We implemented an STL approach and tested the proposed model on tensorial fMRI and matricized DTI data sets of HIV neuroimages. The goal of this study was to improve accuracy in predicting whether someone is HIV positive or negative using fMRI and DTI images of that person's brain

We studied a kernel-based method that differs from traditional STL frameworks that normally use linear model. Before our study there was no knowledge of which kernel would be better for a given study. This is why we needed to evaluate different kernel functions for different datasets. It is a common for different data sets to have different properties, and thus different kernel functions had to be evaluated.

Therefore, based on previous studies, we implemented a polynomial kernel on data sets consisting of fMRI tensor and DTI matrices images of the brain for HIV diagnosis. Based on the observed results observed in experiments, which are presented in the evaluation chapter, it can be concluded that a polynomial kernel performs better on HIV DTI matric data than on tensorial HIV fMRI data. Additional research is needed to determine the most effective kernel learning for fMRI tensorial data.

CHAPTER 8

FUTURE WORKS

Throughout this document, we proposed an extension to general platform DuSK framework by implementation of polynomial kernels. In this chapter, we will go over each contributions that has been made in this thesis and discuss how we can relax some of these constraints or expand some of the models.

- **Integrate decomposition and classification as a whole**

  Similar to what we did in this this thesis, decomposition and classification are usually handle seperately. But, (Wu et al., 2013) developed a supervised tensor-based factorization called Supervised Non-negative Tensor Factorization with Maximum-Margin ($SNTFM^2$) that integrated decomposition and classification and try to handle them as a whole procedure.

  This factorization extended conventional NTF into a supervised decomposition by applying a maximum-margin method (specifically, a support vector machine). This method was developed by coupling the estimation of tensorial data (through accurate reconstruction of tensorial data using additive combinations of the basis) with a maximum-margin constraint (through a generalized discriminative power). According to (Wu et al., 2013) results, $SNTFM^2$ can perform as good as or proposed method or even better depends on the nature of the data set.

- **Use different kernels**

  As previously mentioned in conclusion part, choice of optimum kernel only depends on the nature of data set. For instance, RBF kernel demonstrated a better performance in fMRI data set while polynomial performed better in DTI. In order to improve the result, we can try new kernels or a combination of kernels to boost the learning based on data set

- **Try this method on a different data set**

  As mentioned above, the nature of data set can have a key role in the result of teasing our algorithm performance. Therefore, we can try to run this algorithm on new data sets for example ADNI or ADHD to achieve a wider perspective of the efficiency of the our proposed algorithm.

APPENDIX

AUTHOR'S BIOGRAPHY

She got her B.Sc. degree in Electrical and Computer Engineering with specialization in Electronics from Shahid Beheshti University in 2014. She have done her bachelor thesis on " Action Potential Simulation Therapy (APS) Effect on Relieving pain".

Working on her bachelor thesis introduced her to signal processing and its application on real world problems. So she decided to come to United State and peruse her studies on signal processing areas. She then got admission from University of Illinois at Chicago (UIC) and started her masters from Fall 2015 at UIC. She started doing research with Prof. Mojtaba Soltanalian from summer 2016 on "Signal Design for Eye Tracking with Application to Parkinson Disease Diagnosis and Treatment" and submitted the a paper on Global SIP conference.

On Fall 2016 taking machine learning along with being teaching assistant fro data science course introduced her to magical world of machine learning and its application. Thus, she decided pursue her PhD in computer science.

Although she completed eight hours of her thesis with prof. Mojtaba Soltanalian in order to get more involved in her future major, she decided to start a computer science related thesis with distinguished professor in computer science, professor Philip S. Yu. This document is the result of her research under Professor Yu's supervision.

CITED LITERATURE

[Andersen and Rayens, 2004] Andersen, A. H. and Rayens, W. S.: Structure-seeking multilinear methods for the analysis of fmri data. NeuroImage, 22(2):728–739, 2004.

[Bourennane et al. , 2010] Bourennane, S., Fossati, C., and Cailly, A.: Improvement of classification for hyperspectral images based on tensor modeling. IEEE Geoscience and Remote Sensing Letters, 7(4):801–805, 2010.

[Cai et al. , 2006] Cai, D., He, X., and Han, J.: Learning with tensor representation. Technical report, 2006.

[Carroll and Chang, 1970] Carroll, J. D. and Chang, J.-J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. Psychometrika, 35(3):283–319, 1970.

[Chong and Zak, 2013] Chong, E. K. and Zak, S. H.: An introduction to optimization, volume 76. John Wiley & Sons, 2013.

[De Lathauwer, 1997] De Lathauwer, L.: Signal processing based on multilinear algebra. Katholieke Universiteit Leuven, 1997.

[Geng et al. , 2011] Geng, X., Smith-Miles, K., Zhou, Z.-H., and Wang, L.: Face image modeling by multilinear subspace analysis with missing values. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(3):881–892, 2011.

[Hao et al. , 2013] Hao, Z., He, L., Chen, B., and Yang, X.: A linear support higher-order tensor machine for classification. IEEE Transactions on Image Processing, 22(7):2911–2920, 2013.

[Hardoon and Shawe-Taylor, 2010] Hardoon, D. R. and Shawe-Taylor, J.: Decomposing the tensor kernel support vector machine for neuroscience data with structured labels. Machine Learning, 79(1-2):29–46, 2010.

[Harshman, 1970] Harshman, R. A.: Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis. 1970.

[Håstad, 1990] Håstad, J.: Tensor rank is np-complete. Journal of Algorithms, 11(4):644–654, 1990.

[He et al. , 2014] He, L., Kong, X., Yu, P. S., Yang, X., Ragin, A. B., and Hao, Z.: Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. In Proceedings of the 2014 SIAM International Conference on Data Mining, pages 127–135. SIAM, 2014.

[Kolda and Bader, 2009] Kolda, T. G. and Bader, B. W.: Tensor decompositions and applications. SIAM review, 51(3):455–500, 2009.

[Kong and Yu, 2014] Kong, X. and Yu, P. S.: Brain network analysis: a data mining perspective. ACM SIGKDD Explorations Newsletter, 15(2):30–38, 2014.

[Kong et al. , 2013] Kong, X., Yu, P. S., Wang, X., and Ragin, A. B.: Discriminative feature selection for uncertain graph classification. In Proceedings of the 2013 SIAM International Conference on Data Mining, pages 82–93. SIAM, 2013.

[Kotsia and Patras, 2011] Kotsia, I. and Patras, I.: Support tucker machines. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 633–640. IEEE, 2011.

[Li et al. , 2006] Li, J., Allinson, N., Tao, D., and Li, X.: Multitraining support vector machine for image retrieval. IEEE Transactions on Image Processing, 15(11):3597–3601, 2006.

[Liu et al. , 2008] Liu, Y., Wu, F., Zhuang, Y., and Xiao, J.: Active post-refined multimodality video semantic concept detection with tensor representation. In Proceedings of the 16th ACM international conference on Multimedia, pages 91–100. ACM, 2008.

[Lu et al. , 2008] Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N.: Mpca: Multilinear principal component analysis of tensor objects. IEEE Transactions on Neural Networks, 19(1):18–39, 2008.

[Mocks, 1988] Mocks, J.: Topographic components model for event-related potentials and some biophysical considerations. IEEE transactions on biomedical engineering, 35(6):482–484, 1988.

[Pirsiavash et al. , 2009] Pirsiavash, H., Ramanan, D., and Fowlkes, C. C.: Bilinear classifiers for visual recognition. In Advances in neural information processing systems, pages 1482–1490, 2009.

[Savicky and Vomlel, 2007] Savicky, P. and Vomlel, J.: Exploiting tensor rank-one decomposition in probabilistic inference. Kybernetika, 43(5):747–764, 2007.

[Shashua and Levin, 2001] Shashua, A. and Levin, A.: Linear image coding for regression and classification using the tensor-rank principle. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 1, pages I–I. IEEE, 2001.

[Tao et al. , 2005] Tao, D., Li, X., Hu, W., Maybank, S., and Wu, X.: Supervised tensor learning. In Data Mining, Fifth IEEE International Conference on, pages 8–pp. IEEE, 2005.

[Vapnik, 2013] Vapnik, V.: The nature of statistical learning theory. Springer science & business media, 2013.

[Wang et al. , 2011] Wang, X., Foryt, P., Ochs, R., Chung, J.-H., Wu, Y., Parrish, T., and Ragin, A. B.: Abnormalities in resting-state functional connectivity in early human immunodeficiency virus infection. Brain connectivity, 1(3):207–217, 2011.

[Wolf et al. , 2007] Wolf, L., Jhuang, H., and Hazan, T.: Modeling appearances with low-rank svm. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–6. IEEE, 2007.

[Wu et al. , 2013] Wu, F., Tan, X., Yang, Y., Tao, D., Tang, S., and Zhuang, Y.: Supervised nonnegative tensor factorization with maximum-margin constraint. In AAAI, 2013.

[Yada Zhu and Lawrence, 2012] Yada Zhu, J. H. and Lawrence, R.: Hierarchical modeling with tensor inputs. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pages 1233–1239, 2012.

VITA

Neshat Mohammadi

Education	MS.c. Electrical and Computer Engineering University of Illinois at Chicago	2015 – 2017
	B.S. Electrical Engineering (Electronics) Shahid Beheshti University	2009 – 2014
Research Experiences	Supervised Tensor Learning with Applications, Spring 2017. M.Sc. Thesis, University of Illinois at Chicago. Adviser: UIC Distinguished Professor and Wexler Chair in Information Technology Philip S. Yu. Working on "Tensor Decomposition ", "Higher Order Tensor Decomposition Application", and "Supervised Tensor Learning."	
	Signal Design for Eye Tracking with Application to Parkinson Disease Diagnosis and Treatment, Fall 2017. Research Collaboration with Wave Optimization Laboratory (Wave Opt) Lab, University of Illinois at Chicago. Adviser: Assistant Professor Mojtaba Soltanalian. Completed "Signal Optimization " and " Non- Linear Optimization." Paper Submitted to GlobalSIP Conference.	
	New Business Ranking Prediction Prepared for Yelp Competition, Fall 2016. Course Project, University of Illinois at Chicago. Adviser: Assistant Professor Brian Ziebart. Completed "Ranking Prediction", "Training the Predictive model", and "Finding Optimum Predictor."	
	Simulation and Design of Magnetic Levitation System with Optical Sensors, Spring 2014. Course Project, Shahid Beheshti University. Adviser: Associate Professor Alireza Rezazade. Completed "Stability of Mag Lev System", "Implementation of Current Feedback," and "Optimize System Using Sliding Method with Presence of Noise."	

Study Action Potential Simulation (APS) Therapy effect in relieving pain, Spring 2013.
B.Sc Thesis, Shahid Beheshti University.
Adviser: Assistant Professor Somayeh Timarchi.
Completed "Signal Design and Bio- Signal Simulation."

Job Experience

Teaching Assistant

Teaching Assistant for "Data Science 1" (Fall 2016) and "Data Science 2" (Spring 2017)
Supervised by: Adjunct Professor Ashkan Sharabiani - University of Illinois at Chicago.

Volunteer Jobs

Volunteer for "Intro to Ruby" event held by Anita Borg Institute(ABI) in Chicago, Summer 2017.

Volunteer member of Society of Women in Engineering (SWE) at University of Illinois at Chicago, Fall 2017.

Professional Positions Offer received from Zebra Technologies Corporation for Spring 2018, Illinois, USA.

Enghelab Physical Therapy Clinic at Summer 2014, Tehran, Iran.

Tehran Regional Electricity Company at Summer 2013, Tehran, Iran.

Skills

Programming language: MALTAB; C/C++; SQL; Ruby; Python; Assembly.

Software: Proficient in LaTeX, Microsoft Office; Familiar with Altium Designer, PHP, HTML.

Hardware: Microprocessors, Logic Circuits, Oscilloscopes.

Languages: Proficient in English, Native Speaker in Persian, Intermediate in French and Arabic, basic in Spanish.

Awards and Honors

Chicago Coder Conference Registration Award by ABI Chicago, Summer 2017.

Nominated for Grad SWE board at University of Illinois at Chicago, Spring 2017.

Member of Graduate Student Council (GSC) at University of Illinois at Chicago, Fall 2017.

Getting internship offer from the biggest electrical company of my country passing a highly competitive application process, Tehran, Summer 2013.