学校编码：10384　　分类号_____密级_____
学号：15420141152004　　UDC_____

厦门大学

硕 士 学 位 论 文

复杂关联数据的张量模型与应用研究

Tensor Model of Complex Correlated Data and Application Research

王 丹

指导教师姓名：钱争鸣 教授

专 业 名 称：统计学

论文提交日期：2017 年 4 月

论文答辩时间：2017 年 4 月

学位授予日期：2017 年 6 月

答辩委员会主席：_____

评 阅 人：_____

2017 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2017 年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　年 月 日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2017 年　　月

摘　要

在许多研究领域，数据呈现出多模态的结构特性，张量可以清晰完整地表示这类数据。而张量数据的向量化处理丢失了大量的数据结构信息，甚至造成维数灾和过拟合现象。张量分析的方法直接输入张量数据，能够有效保持数据的结构信息；此外，以张量数据为基础的模型和算法，在减少参变量数目的同时，缓解向量方法在模型学习时极易出现的过拟合现象，这意味着张量模型处理高维小样本问题更加有效，为分析高维向量数据提供了一种新的思路。

基于张量的数据分析方法具有更加广泛的应用前景，所以本文探讨两类复杂关联数据的张量模型及其应用，一类具有天然的张量结构，一类可以转化为张量进行处理。本文的研究内容主要包括以下几个方面：

1.对象关联数据的张量模型及其应用研究。传统方法破坏了张量数据的结构特性，张量可以有效表示该类复杂关联数据。以社会标注系统为背景，用张量及张量分解模型研究该系统的高阶结构特性和统计特性。以系统中的用户、资源、标签三大对象作为三个维度，引入权重值区分“用户-标签-资源”三元关系的关联强度，建立三阶权值张量模型，张量分解后得到最优核张量和三个维度的特征矩阵，以及新的三元关系的评分值，根据最优的评分值产生推荐列表，向用户推荐资源或标签。

2.属性关联数据（向量数据）的张量模型及其应用研究。在实际问题中，存在很多具有关联关系的属性数据，通常采用向量方法进行处理，向量维数过高时容易产生过拟合现象，张量分析方法在减少模型参变量的同时缓解或避免了过拟合现象。本文基于向量数据的学习方法，主要是支持向量机的分类、回归和特征选择方法，自然地推广到张量空间，得到支持张量机的分类模型、回归模型和特征选择方法，实验对比证明，支持张量机模型不仅可以分析向量数据，还可以有效缓解高维小样本问题。

3.实证研究：基于张量空间模型的网络舆情分析。通过张量方法向用户个性化推荐网络舆情热点话题，并实现网络舆情文本的有效识别和自动分类。在实现个性化推荐时，用张量对用户和网络舆情建模，通过张量分解分析用户的兴趣倾向，进而向用户个性化推荐网络舆情热点话题，实验结果表明，引入权重值的张

量分解模型将进一步提高推荐资源的准确率，使个性化推荐结果更加精确。在实现网络舆情文本的有效识别时，将文本表示为20×20的二阶张量，构造张量分类器对网络舆情文本进行分类，实验结果表明，支持张量机模型在解决网络舆情文本的高维小样本问题和数据偏斜问题时具有更好的泛化性能。可见，将张量空间模型应用于网络舆情分析领域具有广泛的应用价值。

关键词：多模态数据；高维数据；张量；张量分析

Abstract

In many areas of research, data presents multimodal structure property, this type of data can be expressed clearly and completely by tensor. Vectorization of tensor data lose many data structure information, even result in curse of dimensionality and over-fitting. While tensor analysis methods enter tensor data directly, keeping data structure information effectively; models and algorithms based on tensor data, reduce parameter variables, ease overfitting phenomenon of vector methods in model learning process, this means tensor model is more effective for high dimension small sample problem, provide a new idea for analyzing high dimension vector data.

Data analysis methods based on tensor have wider applications, so, this paper will discuss tensor models of two types of complex correlated data and their applications, one has natural tensor structure, another can be converted to tensor for processing. Research mainly includes following aspects:

1. Tensor model of objects correlated data and application research. Traditional methods destroy structural property of tensor data; tensor can effectively express this type of complex correlated data. Taking social tagging system as background, using tensor and tensor decomposition model to study higher order structural and statistical characteristics of the system. Taking user, resource, tag the three objects as three dimensions, introducing weights to differentiate "user-label-resource" the ternary relation intensity, establishing the three-order weights tensor model, we can get optimal kernel tensor、feature matrix of three dimensions and rating scores of new ternary relations by tensor decomposition, get recommendation list by rating scores, and recommend resources or tags to users.

2. Tensor models of attributes correlated data (vector data) and application research. In realistic problems, there are many attribute data which has associated relations, often handled by vector methods, it will appear overfitting phenomenon when vector dimension is too high, tensor analysis methods ease or avoid occurrence of overfitting by reducing model parameter variables. The paper naturally generalizes to tensor space based on learning methods of vector data, primarily the classification,

regression, and feature selection of support vector machine, finally obtain tensor classification model, tensor regression model, and tensor feature selection methods, example comparisons prove that support tensor machine model not only can be used to analyze vector data, but also can alleviate the high dimensional small sample problems effectively.

3. Empirical study: network public opinion analysis based on tensor space model. By tensor methods, we personalized recommend hot topics of network public opinion to users, and realize the effective identification and automatic classification of network public opinion texts. In the realization of personalized recommendation, using tensor to establish model for user and network public opinion, analyzing interest tendency of user by tensor decomposition, and then recommending hot topics of network public opinion to users individually, experiment results show that introducing weights will increase accuracy of resource recommendation and make personalized recommendation more precise. In the realization of effective identification of network public opinion texts, text is expressed by two order tensor of 20×20, constructing tensor classifier to classify network public opinion texts, experiment results show that support tensor machine model has better performance in solving high-dimensional small sample problem and data skew problem of network public opinion texts. It can be seen that tensor space model has wide application value in the field of network public opinion analysis.

Keywords: Multimodal data；High dimensional data；Tensor；Tensor analysis.

目录

Contents

第一章 绪论

1.1 研究背景及意义

随着信息科技和互联网技术的快速发展，我们进入大数据时代。大数据时代的到来使传统的数据处理方式开始面临严峻挑战。大数据不仅指数据的数据量庞大，还包括多模态、多属性的复杂关联数据。科学、网络及社会领域中许多数据呈现出多模态的结构特性，如网络流量、社会关系网、彩色图像等，它们的数据结构特性是高阶的，即张量特性，或者可以被组织成张量结构，张量是对这类数据自然而本质的表达方式，可以保留数据元素相互间的内在关联和空间拓扑结构。此外，张量作为向量和矩阵的推广，具有良好的计算特性和分析表达能力。所以，张量是处理多模态高维数据的有力工具。

现实世界中这些具有高阶特性的数据用张量这种数据结构可以直观精确地表达，如果想要进一步分析这类数据，传统方法是将其转化为向量形式，用基于向量的机器学习方法进行分析。对于张量数据的直接转化将导致高维向量的产生，在学习过程中容易出现维数灾和过拟合现象，而且当转换为向量时，原始数据的高阶结构特性遭到破坏，使数据包含的一些重要信息也不复存在。因此，为了应对处理这类高阶张量数据时遇到的问题，许多研究学者深入探索了基于张量数据的机器学习方法，这些方法对张量数据并不作处理，而是将其直接输入，这样，这些数据的高阶结构特性得到了保持，除此之外，张量算法还具有一定的特殊性，直接输入张量数据训练张量模型时需要确定的参变量数目会减少，缓解或者避免了过拟合现象。

除了多模态结构的张量数据外，现实中更多的是包含多个属性的向量数据。在实际问题中，一个对象往往包含多个属性，举一个简单的例子，人的属性包括其名字、年龄、性别、人际关系等，商品的属性包括其名称、类别、材质、特殊说明等。处理这类数据的方法一般是基于向量的统计学习方法，但是大数据时代数据量爆炸式的增长，使向量数据的维数呈几何级数式增长，而在现实研究中，存在很多我们人为无法控制和解决的因素，如样本的获取方式、成本以及一些不