《大数据分析及应用》期中报告要求

**主题：**

从以下主题中任选一个，进行深入调研和学习，撰写期中报告。形式可以采用以下两种：

（1）针对**某一算法**的设计和实现进行深入调研和分析，论述该算法的设计思想、实现步骤（实现过程以及具体算法流程）、主要算法（算法伪代码或者算法源码）、具体实验分析。***主要阐述某一算法的改进和优化进展。***

（2）针对**某一个问题的多种相关算法**进行深入调研和分析，论述不同算法的设计思想，算法流程，分析不同算法的各自优点和缺点，实验结果对比分析。***主要阐述针对某一问题的多种解决方法的特点。***

选题1：基于社交网络数据的抑郁症用户画像

选题2：基于社交网络数据的抑郁症用户识别

选题3：基于多模态数据的情感分析

选题4：基于POI数据的城市热点发现

选题5：基于出租车GPS数据的交通流量预测

***注意：针对每一个选题，至少有2种算法分析，可以是一类算法及该算法的改进方法，也可以是针对一个问题的不同算法。***

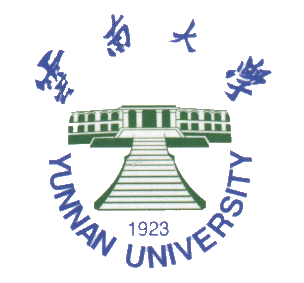
**要求：**

（1）报告格式规范：按科研论文格式撰写，主要包括：标题、作者、中文摘要、关键词、论文主体内容（背景和意义，算法描述，实验分析）、总结、参考文献（至少包含3篇近三年的科研论文）。参考附件1。

（2）报告不少于3000字。

（3）报告提交时间：2023年11月27日

（4）要求提交报告电子版和纸质版，电子版标题“组长学号-姓名-期中报告”（期中报告2人一组，期末4人一组）

云南大学软件学院期中课程报告

Midterm Course Report

School of Software, Yunnan University

**个人成绩**

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **学号** | **姓名** | **成绩** |
| **1** | **20211060245** | **陈俊宏** |  |
| **2** | **20211120171** | **孔令高** |  |
| **3** |  |  |  |
| **4** |  |  |  |

学　　期: 　2023秋季学期

课程名称: 　大数据分析及应用

任课教师: 　何婧

报告题目: 多模态的情感分析

组长姓名:

联系电话:

电子邮件:

完成提交时间：2023年 月 日

作业截止时间：2023年11月27日

**小组成员及分工**

**（工作量总和为100%，真实体现成员工作量，不能直接用平均值）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **学 号** | **姓 名** | **工作内容** | **工作量（%）** |
| 1 | 20211060245 | 陈俊宏 |  |  |
| 2 | 20211120171 | 孔令高 |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 6 |  |  |  |  |

评分标准：

1. 选题。 10%

2. 综述完整，结构组织合理，内容丰富，文字表达清楚。40%

3. 报告格式规范，图表清晰。20%

4. 有基础算法实验代码及分析。10%

5. 有自己的观点。10%

6. 课堂汇报逻辑清楚，PPT展示完整。10%

**（注：此模板非完整论文，已做删减。只看格式，勿参考内容）**

**异质网中基于张量表示的动态离群点检测方法**题目三号

#### Tensor Representation Based Dynamic Outlier Detection Method in Heterogeneous Network Title四号

***根据选题，自拟题目，题目比选题方向更为具体***

刘露1 左万利1,2 彭涛1,2作者四号宋体，作者和单位的对应关系标注在作者姓名的右上角

1（吉林大学计算机科学与技术学院 长春 130012）单位小五号

2（符号计算与知识工程教育部重点实验室（吉林大学）长春 130012）

1（liulu12@mails.jlu.edu.cn）小五号

摘要 挖掘隐藏在异质信息网络中丰富的语义信息是数据挖掘的重要任务之一．离群点在值、数据分布、和产生机制上都明显不同于正常数据对象．检测离群点并分析其不同的产生机制，最终消除离群点具有重要的现实意义．目前，针对异质信息网络动态离群点检测的研究工作相对较少，还有很多问题有待解决．由于异质信息网络的动态性，随着时间的变化，正常数据对象也可能转变为离群点．针对异质网络提出一种基于张量表示的动态离群点检测方法，并根据张量表示的高阶数据构建张量索引树．通过搜索张量索引树，将特征加入到直接项集和间接项集中．同时，根据基于短文本相关性的聚类方法来判断数据集中的数据对象是否偏离其原聚簇来动态检测网络中的离群点．该模型能够在充分降低时间和空间复杂度的条件下保留异质网络中的语义信息．实验结果表明，该方法能够快速有效地进行异质网络环境下的动态离群点检测．

摘要五号楷体，300字左右

关键词动态离群点检测；异质信息网络；张量表示；张量索引树；聚类 关键词五号楷体

中图法分类号 TP391 分类号五号

正文五号宋体

## 0 引言 一级标题小四黑

异质信息网络代表一个现实世界的抽象，专注于多种类型的对象以及对象之间的相互关系．异质网络中经常存在许多不同于正常对象的离群点．作为数据挖掘领域的一个重要分支，离群点检测可以预测数据对象行为和发展趋势，具有很重要的现实意义．离群点检测有着广泛的应用，例如，异常天气检测[1]、信用卡欺诈检测[2]、心电图分析[3]、异常GPS追踪[4]、文本挖掘中异常的主题检测[5]等．参考文献顺序标引

……

本文的主要贡献如下:

1) 我们提出了一种基于张量表示的异质网络动态离群点检测方法TRBOutlier，通过分析网络中数据变化趋势判断其是否为离群点;

2) 张量表示方法被应用到异质网络中来处理不同类型的数据，张量索引树的构建解决了数据稀疏性问题，同时保留了数据的语义关系;

3) 在张量索引树的基础上对网络中出现的短文本进行相关性分析，并依据短文本的相关性对异质网络中的实体进行聚类;

4) 应用不同数据集的实验结果表明我们提出的离群点检测算法可以有效发现异质网络中存在的动态离群点．

## 1 相关工作 一级标题小四黑

离群点检测不论在同质网络中还是在异质网络中都发挥着重要的作用．静态离群点和动态离群点检测在不同的背景下也都有着广泛的应用和重要的研究意义．接下来，我们概述已有的部分离群点检测工作以及在不同条件下的应用．

关于离群点检测的研究有很多，但大多数都是针对同质信息网络的研究[12-13]．文献[14]提出了一种基于密度的局部离群点检测算法．该方法通过引入信息熵来发现网络中存在的局部离群点．文献[15]提出了一种使用后缀树的离群点检测方法．该方法认为离群点稀少，出现的次数也相对较少，比一些周期性出现且出现次数频繁的正常点更加重要．其主要用于处理数值或者字符，因此被应用于同质信息网络．

……

本文提出的基于张量表示的动态离群点检测方法，将网络中的异质数据进行动态分析．不仅解决了数据的稀疏性问题，也很大程度上保留了数据之间的语义关系．该方法可以根据异质数据所在聚簇是否发生变化来判断网络中的数据是否为离群数据，也可以根据离群数据来分析其产生机制并进行相应的处理．

**2**异质信息网络中的张量表示方法

在本节中，我们主要介绍张量表示方法在异质信息网络中的应用，并且将类型的概念引入张量中．其中，2.1节给出了模型所需的定义．2.2节描述了如何对异质数据来构建张量索引树．

**2. 1**基本定义 二级标题五号黑

对同质网络进行离群点检测时， 通常用数值或向量来表示网络中的实体．例如，在异常天气检测中，气温用数值进行记录，1周或1个月的气温值可以存储在1个向量之中．在对文本主题异常检测时，文本中的特征权值通常用向量表示．不论气温还是文本，数值中和向量中存储的都是同一类别的实体，即数值和向量的定义域是相同的．然而，在异质信息网络中存在着不同类型的实体和链接，将所有实体各自表示成向量进行相似或离群的计算往往不能得到满意的结果．因此，在本节中，我们提出了一种张量表示方法来处理异质网络中的实体．将张量矢量化处理很可能引起维数灾难并破坏了原本高维数据之间的结构关系．……将张量表示用于异质离群点检测是一种新的尝试．在详细描述离群点检测算法之前，我们先给出一些基本的符号解释和定义．

定义1．异质信息网络[24]．给定一个有向图 *G*=(*V*，*E*;;*A*，*R*)．*V*代表节点集，*E* 代表边集．表示对象类型映射函数．表示关系类型映射函数．(*v*)∈*A*表示每个对象*v*∈*V*都属于一个特定的对象类．(*e*)∈*R*表示每个关系*e*∈*E*都属于一种特定的关系类．当节点类型数量∣*A*∣>1或边的类型数量∣*R*∣>1时，这样的信息网络被称为异质信息网络．反之为同质信息网络．

定义2．异质信息网络中的*N*阶张量．我们将异质信息网络中的*N*阶张量表示为***X***∈．其中， *N*为张量的阶数(也可称模数)．第1阶张量的长度记为*l*1，其对应的实体类型为*type*1．第*N*阶张量的长度记为*lN*，其对应的实体类型为*typeN*．*typei*∈*A*， 1≤*i*≤*N*．

公式用office2003编辑器或者mathtype，普通变量用斜体，向量、矢量、张量用黑斜体

例如，文献网络可以被表示成1个4阶张量，***X***∈．电影网络可以被表示为1个4阶张量***X***∈或一个3阶张量***X***∈．

定义3．张量索引树．张量索引树是包含*n*(*n*≥1) 个节点且满足下列条件的有限集合:

1)张量索引树中的每一个节点都由1个 *N*(*N*≥1)维向量组成．

2)存在唯一一个向量节点*X*0，它是由一系列时序数据组成的向量，称为张量索引树的根节点．

3)张量索引树同一层的节点可以包含不同类型的数据．但同一类型的数据都处于张量索引树的同一层中．

**……**

**2. 2**张量索引树的构建

我们之前提到过，张量矢量化可能破坏各阶中数据之间的关系．在大数据环境下，不可避免地会引起维数灾难．而对于异质网络中的数据，如何进行高阶统计也是我们需要考虑的问题．应用张量分解方法，例如构建邻接矩阵，处理数据集中的样本会出现严重的数据稀疏性问题．在本节中，我们提出了一种构建张量索引树的方法来处理异质信息网络中用张量表示的数据，进而动态发现网络中存在的离群点．

我们首先描述张量索引树的构建过程．由于本文主要解决动态离群点发现的问题，因此，张量索引树的根节点为时序数据组成的向量***X***=(*X*1，*X*2，…， *Xn*) ，如图1中Layer 1所示．在本节中，我们使用文献网络中的3阶张量模型来说明构建过程．在时间点*X*1中，顺序遍历*N*(*N*=3)阶张量中的第1阶张量．将属于*type*1类型的实体(作者名)加入第2层向量节点中．同时，将该作者发表的论文依次加入第3层向量节点中．第3层向量节点中的论文题目作为第2层向量节点中作者名的后继节点．……

张量索引树的构建不仅可以加快离群点检测的速度，而且可以有效解决在使用邻接矩阵进行计算时出现的数据稀疏性问题．



Fig. 1 The tensor index tree constructed by a third order tensorin bibliographic network.

图1 文献网络中由3阶张量所构建的张量索引树

中英文图题用小五号，图内的字用英文六号

**3**基于张量表示的聚类过程

在本节中，我们详细介绍如何根据给定的入口*entry*(*source*，*target*)搜索张量索引树，进而使用聚类方法发现异质网络中存在的离群点(即源节点相对于目标类别是否离群)．张量索引树可以根据给定入口快速定位相关的异质信息．……

首先，提取索引树中源节点至目标类别路径上直接关联的数据，将其放入直接项集*DIS*(direct item set)中，如图2表格中的Paper1~Paper3．根据频繁项集的思想，将包含直接项集中目标类型节点2项以上的前驱节点及其相应路径上的直接关联信息加入到间接项集*IIS*(indirect item set)中，如图2表格中的Paper4~Paper10．从表格中可以看到，Paper4中的关键词outlier detection和heterogeneous network在Paper1~Paper3中出现过，因此，Paper4被加入间接项集中．得到直接项集和间接项集后，项集中目标类别的特征被加入到1维向量词典中，记为***Dic***1=(*st*1:*stw*1,*st*2:*stw*2,…,*stN*:*stwN*)．其中，*st*表示短文本特征，*stw* 表示短文本特征权值．式(1)中给出了短文本权值的定义．

, (1)

其中，*NDIS*和*NIIS*分别为直接项集和间接项集中项的数量(1条路径中的数据集合称为1条项集，也称记录)．*tki*表示特征*i*在直接项集记录*k*中出现的次数． *tli*表示特征*i*在间接项集记录*l*中出现的次数．为调节因子．本文中，被设置为0.5．由于直接项集中的记录和源节点的相关程度将高于间接项集中的记录与源节点的相关程度，因此，调节因子起到调节特征在直接项集和间接项集中重要程度的作用．

根据上述特征权值计算过程，我们可以得到源节点相对于目标类别的特征表示，根据特征相似度判断对象是否属于1个聚簇当中．但是，不同的源节点对应的目标类别节点大不相同，需在聚类之前解决数据分布不一致的问题．首先构建1个包含目标类别所有特征的词典*DIC*，每个源节点关于目标节点的特征表示为***Dic****i*，1≤*i*≤*N*sum其中，*N*sum为目标类别特征总数．每个***Dic****i*中的特征均为*DIC*的一部分．计算短文本特征权值的算法在算法1中给出．由于之前已经得到源节点相对于目标类别节点的特征表示，对于所有*DIC*中的特征，如果未在***Dic****i*出现过，则该特征的权值设为0，该方法可以解决聚类过程中出现的数据分布不一致问题．



Fig. 2 Direct item set and indirect item set.

图2 直接项集与间接项集图字六号英文

……我们提出一种基于张量表示的短文本聚类的方法来进行离群点的检测．式(2)给出了2个特征向量进行相似度计算的方法[25]．

, (2)

其中，向量(*stwi*1,*stwi*2,…,*stwiN*sum)和向量(*stwj*1,*stwj*2,…,*stwjN*sum)分别是源节点*i*和 *j*的*N*sum-维特征表示，代表***Dic****i*和***Dic****j*．*Sim*(***Dic****i*，***Dic****j*) 的取值范围为[0,1]．式(3)定义了一个0，1变量*CV* (comparison variation)[25]判断2个聚类能否合并或分裂．

 (3)

其中，*Ci*和*Cj*分别代表2个聚类．***Dic****k* (1≤*k*≤*R*)表示所有属于聚类*Ci*中的特征向量．根据式(3)，若聚类*Ci*中的多数对象都和*Cj*的聚类中心特征相似，那么， 1-*Sim*(***Dic****k*,*Cj*)的数值就会很小，使得max{*Sim*(***Dic****k*,*Cj*)|***Dic****k*∈*Ci*}远大于(1-*Sim*(***Dic****k*, ．此时，应将2个聚簇合并．作为一个调节相似程度严格与否的因子也起到至关重要的作用．

算法1．短文本特征权值计算算法．

输入：张量索引树*TI-tree*， 源节点*s*， 目标类型*t*， 直接项集*DIS*， 间接项集*IIS*；

输出：源节点相对于目标类型的特征表示***Dic****s*．

① *BreadthFirstSearch*(*TI-tree*, *s*)；

② for each path containing *s*

③ *DISgetItemSet*(*s*,*t*)；/\* 将源节点*s*关于目标类型*t*路径中的节点放入集合*DIS*中\*/

④ end for

⑤ *CgetChildNode*(*DIS*,*t*)；/\*将*DIS*中类型*t*

的节点放入集合*C\*/*注释不用双斜线

⑥ if any node *p* is the parent node of at least two

nodes in *C*

……

⑧ *DgetChildNode*(*IIS*,*t*)；/\*将*IIS*中类型*t*的节点放入集合*D\*/*

⑨ end if

……；

**4**基于张量表示的离群点检测方法

离群点和正常数据对象的产生机制不同，以致于出现了不同于正常点的数值或数据分布等现象．……在本节中，我们提出了一种基于张量表示的动态离群点检测方法．

……

在真实的网络环境中，我们可以将数据分为*N*个时刻进行数据分析．……图4给出了基于张量表示的动态离群点检测方法的整体框架．



Fig. 3 Changes of outliers from *time*1 to *tim*e2.

图3 时刻*time*1到时刻*tim*e2基于聚类方法的离群点变化模型



Fig. 4 The overall framework of dynamic outlier detection method.

图4 基于张量表示的动态离群点检测方法的整体框架

**5** 实验与结果

在本节中，我们使用本文提出的技术构建了一个面向异质网络的动态离群点检测模型，并且在2个数据集(AMiner[10]和Yahoo!Movies[11])上测试了我们的方法．

**5. 1** 度量标准

离群点的判断标准一般分为2种：……我们定义精确率和召回率公式如下．

, (4)

, (5)

其中，*ChS*表示聚类过程中所在聚簇发生变化的数据集．*OutS*表示聚类前被标记为离群点的数据集．

……

**5. 2** 数据集

本文将在AMiner和Yahoo!Movies 2个数据集上测试我们提出的方法．下面我们从数据集大小和数据特点等方面分别介绍这2个数据集．

AMiner数据集是一个异质的文献信息网络．它由3个部分构成，分别是AMiner-Author， AMiner-Paper， AMiner-Coauthor．……Yahoo!Movies数据集作为Yahoo!数据集评分和分类数据的一部分，可以被应用到各种数据挖掘算法当中．……在AMiner和Yahoo!Movies 2个数据集中均构建3阶张量进行实验．

**5. 3** 实验和结果

在本节中，我们进行4个实验来验证我们提出的离群点检测模型的可行性和高效性．第1个实验是通过变化的取值判断其取何值时精确率率和召回率可以达到最大值．……

**Table 1 The Results on AMiner**

**表1** **TRBOutlier方法在Aminer上的实验结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *OutPrecision* | *OutRecall* | *F-Measure* | *Accuracy* |
| 0.4 | 1 | 0.32 | 0.485 | 0.454 |
| 0.6 | 1 | 0.385 | 0.556 | 0.527 |
| 0.8 | 0.98 | 0.43 | 0.598 | 0.635 |
| 1 | 0.956 | 0.525 | 0.678 | 0.752 |
| 2 | 0.937 | 0.695 | 0.798 | 0.896 |
| 3 | 0.915 | 0.83 | **0.870** | **0.938**  0.861 |
| 4 | 0.831 | 0.855 | 0.843 |

三线表，表题用小五号黑体,中英文表题，表的内容用英文小五号

在第2个实验中，我们将动态离群点检测算法和2个基线算法进行*F-Measure*和*Accuracy*的比较．……TRBOutlier的离群点检测性能要高于EBC和CDOutlier (如图5所示)．



此图坐标值不规范，最好在0.1~1000内。

正确标为2,4,6……，在标目位置写 10-3×Number of Objects

1. AMiner



此图坐标值不规范，最好在0.1~1000内。

1. Yahoo!Movies

Fig. 5 Performance comparison on AMiner and Yahoo!Movies.

图5 3种离群点检测方法的性能比较

图题用小五号中英文，分图题只用英文，图内容用英文

……

**6** 总结

本文提出了一种基于张量表示的动态离群点检测方法．静态离群点检测方法可以发现1个时间点内的异常数据，然而网络中的数据是不断变化的．……实验结果表明，本文提出的动态离群点检测方法在准确度和效率方面优于许多已有的离群点检测方法．因此，本文的方法是可行且高效的．

**参 考 文 献**

1. Shepherd J M, Burian S J. Detection of urban-induced rainfall anomalies in a major coastal city[J]. Earth Interactions, 2003, 7(4): 1-17期刊文献要有年、卷、期、起止页码（或编号），期刊名称不缩写
2. Beutel A, Faloutsos C. User behavior modeling and fraud detection[J]. IEEE Intelligent Systems, 2016, 31(2):84-86
3. Jiang B C, Yang W H, Yang C Y. An SPC-based forward-backward algorithm for arrhythmic beat detection and classification[J]. Industrial Engineeering & Management Systems, 2013, 12(4): 380-388
4. Chen C, Zhang D, Castro P S, et al. Real-time detection of anomalous taxi trajectories from GPS traces[M]// Mobile and Ubiquitous Systems: Computing, Networking, and Services. Berlin: Springer, 2011: 63-74

图书文献要有出版地、出版社、年、起止页码（或不写）。出版地指出版社所在的城市

1. Srivastava A, Zane-Ulman B. Discovering recurring anomalies in text reports regarding complex space systems[C]//Proc of IEEE Aerospace Conf. Piscataway, NJ: IEEE, 2005: 55-63

会议文献要有论文集名称、出版地、出版社、年、起止页码。注意：出版地是出版社所在的城市，不是会议开会地点。网络下载需要提供下载日期和下载地址

1. Lebedev L P, Cloud M J, Eremeyev V A. Tensor Analysis with Applications in Mechanics[M]. Hackensack, NJ: World Scientific, 2010
2. Schobeiri M T. Vector and tensor analysis, applications to fluid mechanics[M]// Fluid Mechanics for Engineers. Berlin: Springer, 2010: 11-29
3. Hao N, Kilmer M E, Braman K, et al. Facial recognition using tensor-tensor decompositions[J]. SIAM Journal on Imaging Sciences, 2013, 6(1): 437-463
4. Liţă L, Pelican E. A low-rank tensor-based algorithm for face recognition[J]. Applied Mathematical Modelling, 2015, 39(3): 1266-1274
5. Tang J, Zhang J, Yao L, et al. Arnetminer: Extraction and mining of academic social networks[C]//Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990-998
6. Yahoo! Webscope Program. Yahoo! Movies user ratings and descriptive content information, v1.0 [OL]. [2016-01-28] http://webscope.sandbox. yahoo.com

网络文献要有下载日期和下载地址）

1. Wu S, Wang S. Information-theoretic outlier detection for large-scale categorical data[J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25(3): 589-602
2. Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: A survey[J]. Data Mining and Knowledge Discovery, 2015, 29(3): 626-688
3. 胡彩平, 秦小麟. 一种基于密度的局部离群点检测算法DLOF[J]. 计算机研究与发展, 2010, 47(12): 2110-2116