非负矩阵分解在交通数据挖掘上的应用

【摘要】非负矩阵分解为特征抽取提供了一种崭新的方法。由于对分解结果加入非负限制，基于非负矩阵分解抽取的特征向量更能反映样本的局部特征，因而更接近人们的认知习惯，并具有很高的可解释性和预测性。本文将非负矩阵分解运用于交通数据挖掘，研究其在信息获取领域中的应用，设计相应的处理算法，提高处理效果。

本文首先全面介绍现有的非负矩阵分解基本思想、基本算法。然后从理论上将非负矩阵分解算法与主成分分析、奇异值分解等常用特征抽取方法进行比较，通过比较总结出基于非负矩阵分解所抽取的特征向量具有局部性和可解释性等特点。在理论分析基础上，再将非负矩阵分解应用于信息获取领域，研究基于非负矩阵分解的特征抽取方法在处理交通数据挖掘的应用。

本文的算法设计是对车辆进行建模，把每量车辆全天的行驶情况抽象为一种交通模式，用一个 24 维向量来代表一辆车的模式，用 NMF 方法在若干次实验后，得到了的车辆出行的基本的三种特征：上下班特征、公司间往返（上班时段）特征、其他（主要是夜晚）特征。实际上，这种车辆特征就是车辆行驶的时间特征。进一步，本文再用基本的车辆出行基本模式去拟合一个路口的交通模式。 路口交通模式的区别其实可以体现为一天中车辆主要集中出现的时间段不同。经过最小二乘拟合，也得到了不同的三种路口。

【关键词】非负矩阵分解；交通模式；车辆特征

[Abstract] Nonnegative matrix is decomposed into feature extraction provides a new

method. Because join nonnegative constraints, the result of the decomposition based on nonnegative matrix decomposition to extract the feature vector can reflect the local characteristics of samples, and therefore closer to people's cognitive habits, and highly interpretability and predictability. This article will nonnegative matrix decomposition is applied to traffic data mining, and studies its application in the field of access to information, design the corresponding processing algorithms, improve the treatment effect.

This paper first introduces the existing basic ideas, basic nonnegative matrix decomposition algorithm. Nonnegative matrix decomposition algorithm and then from theory and principal component analysis, singular value decomposition (commonly used feature extraction methods such as comparison, by comparing the summed up based on nonnegative matrix decomposition feature vector extraction by has the characteristics of locality and interpretability. On the basis of theoretical analysis, and then nonnegative matrix decomposition is applied to the field of information retrieval, the study of feature extraction method based on nonnegative matrix decomposition in dealing with traffic applications of data mining.

Algorithm design of this paper is carried out on the vehicle model, the amount per vehicles throughout the day is abstracted as a transport mode, with a 24 d vector to represent a car model, using NMF method after several experiments, the vehicles travel the three basic characteristics: commuting characteristics, company (time) to work back and forth between features, other (mainly) at night. In fact, this kind of vehicles' characteristic is the time characteristics of the vehicle. Further, this paper again with basic vehicle travel mode to fitting a intersection traffic patterns. For the difference intersection traffic patterns can be embodied in a day of vehicles are mainly concentrated in different time period. Through least square fitting, also got different three types of intersections.

[Key Words] Non-negative Matrix Factorziation; traffic pattern; Vehicle characteristics

目录

第一章 引言

矩阵分解在很多领域获得了广泛的应用,在数值代数中，利用矩阵分解可以将规模较大的复杂问题转化为小规模的简单子问题来求解；在应用统计学领域，通过矩阵分解得到原数据矩阵的低秩逼近，从而可以发现数据的内在结构特征.在机器学习和模式识别的应用中，矩阵的低秩逼近可以大大降低数据特征的维数，节省存储和计算资源[1].

1.1 问题背景

关于城市交通各种问题已经在很早些年就引起了科学家们大多注意。根据X.Jin’s关于出租车的交通数据分析[2]，可以知道有各种各样的方法都可以用于解决关于交通数据分析[3][4][5]。随着十几年前提出NMF（非负矩阵分解）方法到现在，我们发现它有效的解决了很多问题，在人脸识别、文本聚类中都有很大的突破。关键是它的具有实际语义的分解结果是对解决很多现实问题来说具有很重要的作用[6][7]。不过也可以发现NMF该方法用于该领域的应用是相当少的。所以从X.Jin’s的论文中可以发现，再处理分析交通数据的问题中，这样的方法是很新颖的，在知道已有的一些交通数据下，能得出一些其他分析方法角度所不同的分析结果。交通数据是很庞大的，有时候我们自己都不知道，拿着这些数据我们到底能获得一些怎样的有意义的对交通改善或者了解交通状况有帮助的数据。X.Jin’s获得的数据是整个大城市的出租车的打表情况，只要有乘客上下车，就每次记录上车的时间地点，和下车的时间地点。有了这些数据，他们通过NMF方法得到了不同区域交通状况信息。目前，我们获得了一个城市所有拍照设备记录下的不同车辆在不同时间点经过的所有数据，考虑到NMF方法快速且具有实际语义的特点，所以我们决定把这样的方法应用于该数据上，建模看能不能得到一些有实际价值的结果。

第二章 方法介绍

2.1 非负矩阵分解定义

信号或信号处理的许多数据具有非负性的特点，如灰度图像、物质成分含量文章中单词出现的次数或者交通流量数据等。在用线性表示方法处理这类数据时往往要求分解的结果（包括基向量和系数）都是非负的。此时若采用传统的因子分析方法，如主成分分析，因为其结果中含有负数而失去了物理意义，而采用非负（正）矩阵分解方法就可以避免这一点[5]。

非负矩阵分解是一种多变量分析方法． 假设处理 m 个 n 维空间的样本数据， 用V_{m*n}表示，该数据矩阵中各个元素都是非负的，表示为 V≥0。对矩阵V_{m*n}进行线性分解，有

$$V_{m*n} \approx B_{m*r} H_{r*n} \qquad (1)$$

其中B_{m*r}称为基矩阵，为H_{r*n}系数矩阵，NMF 不允许 w 和 H 中有负数元素。正是该约束使得在用部分表示整体的过程中，只有加法的运算而没有减法的运算。这个特性反映了由部分构成整体最直观的感受[8]。

2.2 非负矩阵分解意义

非负矩阵分解方法最早是在人脸识别分析中提出的。其作用类似于主成分分析或者因子分析方法，都是从大量样本中找到组成这些样本的特征成分，或者叫做主成分，然后每个样本就可以由这些特征成分线性组合而成，这样可以比较直观地看出哪些样本主要由哪些特征成分组成。

但是主成分分析、因子分析或者奇异值矩阵分解，这些方法，虽然从各种不同角度得到了他们属于自己的比较优秀的结果，例如奇异值矩阵分解的结果在几何角度的解释是值得认可的。但是他们都有个共通的特点，就是其数值结果往往都是有负数出现的。这是没有实际语义的，或者说很难用实际现象去解释它。

所以非负矩阵最大的优势在于，它的结果一定是非负数，这就很符合实际

需求，因为很多实际问题是不会出现负值的。所以通过 NMF 得到的特征成分，是有实际语义的，而且每个样本不仅仅是由这些特征成分的线性组合，更是一种加性组合，也就是纯加性（系数也是非负，这是 NMF 所要求的）。

其实 NMF 的过程很简单，已知一个大矩阵 V_{m*n}，通过数学分析学的方法近似得到两个小矩阵的乘积，分别为 B_{m*r} 和 H_{r*n}，其中 r<m。r 是根据不同问题而定的，r 即是特征数，通常可以根据分解结果的稳定性来确定 r。V 的每个 m 维列向量即代表一个样本，一共 n 个样本。B 代表基矩阵，其每个 m 为列向量代表一个特征成分，一个 r 个特征成分。H 代表系数矩阵，其 r 维列向量表示某一个样本被这 r 个特征成分线性组合表示的系数，一共 n 个这样的列向量。所以分解成功后，我们就得到了主要的特征成分，表示每个样本的 m 维向量信息压缩成 r 维向量信息，并且该 r 维向量更能体现不同样本的区别。

2.3 非负矩阵分解计算

算法中每次迭代 B 和 H 的新值由当前值乘以某个系数而得到，而这些系数取决于式(1)中的近似程度。按照这样的规则不断地迭代，可以确保 B 和 H 收敛到一个局部最优的矩阵分解。

迭代规则：

$$\mathrm{H}_{ij} = \mathrm{H}_{ij} \frac{(B^T V)_{ij}}{(B^T BH)_{ij}} \qquad (2)$$

$$B_{ij} = B_{ij} \frac{(VH^T)_{ij}}{(BHH^T)_{ij}} \qquad (3)$$

按照上述的规则不断地迭代，直到下式所示评价函数达到局部的最小：

$$\mathrm{F} = \sum_i \sum_j (V_{ij} - (BH)_{ij})^2 \qquad (4)$$

该评价函数定义为 V 与 BH 间的欧式距离，并用它来评价两者近似成都[9]。

第三章 实验过程和分析

3.1 数据准备

3.1.1 数据来源

大城市里很多主干道上都有很多交通照相设备，通常这些拍照设备只要有汽车经过就会拍下他们的车牌号。从交通局那里可以获得这样的数据，有了这些数据，我们就可以知道几乎所有车辆在大部分城市区域的活动情况，在哪个时间点什么车通过了什么地方。所以好好利用这些数据对分析交通状况来说是很有意义的。事实上，已获得的数据是，杭州市7月份整个月的所有交通设备的车牌拍摄数据。

3.1.2 数据表示

根据非负矩阵分解算法，我们可以知道关键是要表达出，什么是向量，什么是矩阵，这样我们才能用这个方法。所以在这里，我的具体做法是，把每辆车的形式情况表示成每一个样本向量。也许有人会反问为什么不把某个交通拍摄设备所代表的那个区域的交通状况来表示成一个向量。事实上，这样做也是可行的不过设备的数量（几百）是远远小于汽车的数量（几百万）。在NMF这个方法过程中，样本太少对于分解结果是非常不利的。

因此，怎样把一辆车的交通形势状况转换成一个向量表示又是进一步需要解决的任务。可以先把注意力集中在一天的数据上，首先统计出现的车总数量n。然后，把一天的时间分成m个时间段。由于每个拍摄设备的具体地点和具体拍摄某车的时间，所以我们可以统计出某辆车在这一天中哪些时间点通过了哪些地方，于是就可以大概得到了在每个时间段里每辆车行驶的距离是多少。这个距离肯定是不够准确的，不过可以大致刻画出一辆车在一天中的大概行驶状况。最后对于每辆车的交通行驶状况，我们就转换成一个m维的数值向量，每一维的数值就是这个时间段内该车行驶的大概距离。如果一共有n辆车，那么就有n个这样的m维向量，也就是样本向量集。我们关心的矩阵就出来了V_{m*n}(n个m维列向量并排在一起).

3.2 车辆特征模式

3.2.1 实验过程

已有数据格式是设备号、车牌号、时间，这样一个数据就相当于一次路口拍摄。

一个文件包含一天中24小时全杭州市所有设备拍摄的数据。选出2012年7月份四个礼拜的周二到周四一共12天的数据。（周二到周四相比周一周五更能体现工作日的特征）

数据预处理，去除掉车牌号未知的数据（可能是拍摄过程中一些技术因素造成）。若干个（设备号、车牌号、时间）这样的数据，经过预处理后得到，车总数C，设备总数D。把一天分为24个时段，一辆车24个时段分别通过的设备数代表这辆车的交通模式，即24×1的向量，通过12天的数据统计得到每辆车的平均值。一共有C辆车，那么就组成一个24*C的矩阵V，V(j,i)表示第i辆车在第j小时中通过设备的数量。

通过分解得到V=W*H，W是24*r矩阵，H是r*C矩阵，这里r未知，由人为规定，表示特征数目。r的最终确定还要看分解效果和可解释性。 W代表的是r个基本特征，H代表系数矩阵。

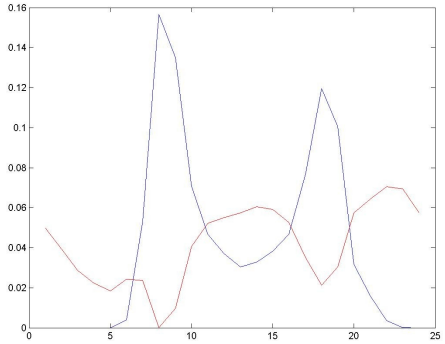
观察到有很多车在12天中出现的次数是相当少的，所以先对车进行分类考察，分为 出现了1～3天的车，出现了4～6天的车，出现了7～9天的车，出现了10～12天的车，分别进行分解，发现1～3天的车稳定性和可解释性极低，4～6天的车可解释性也不高，而后两种则分解效果很好，这表明城市中有相当一部分车辆很少出现甚至可能只是过路车，这样的车并不能反映工作日常规车辆的规律，所以过滤掉出现1～6天的车，再对7～12天的车进行分解。具体的分解方法我是直接用NMF现有的代码，具体数学方法在这里就不做过多的描述。

3.2.2 不同的r值分析

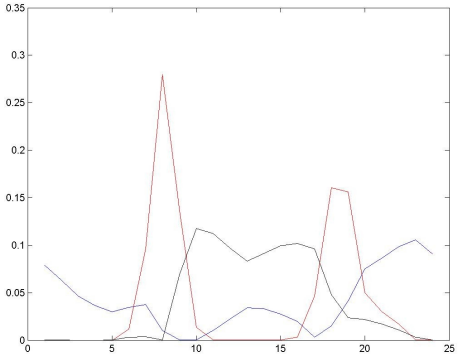
容易想到，在分解过程中，不同的r值能产生不同的分解结果。但是在X. Jin的论文中，我们并没有看到怎样去决定一个合适的r值。相反他们只是说当r=3的时候最稳定。事实上，在我查阅了很多论文发现，并没有一个较好的理论方法可以直接决定r值什么时候是最优的。所以更多的时候，是要结合实际情况，什么r值最适合当下所关心的实际问题，而没有一个统一的方法。考虑到车辆总

数的巨大，可以想到分解结果要得出一个非常稳定的特征模式是比较困难的，因为每辆车的行驶情况都会有一些差别。如果用Cost Function[10]，我们可以知道，r值肯定越大越好。所以我的打算就先不决定r到底为多少，先看分解结果r为多少时更有实际语义解释。

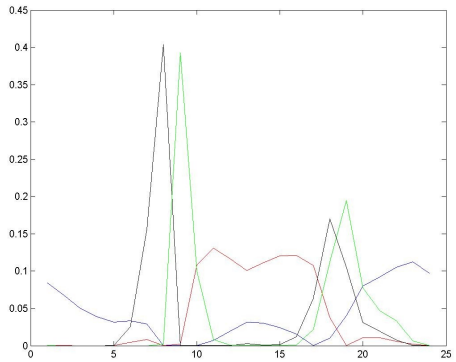
3.2.3 分解结果



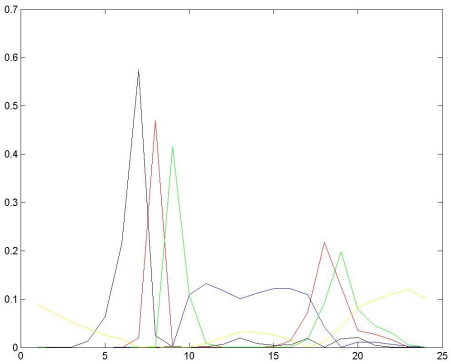
r=2



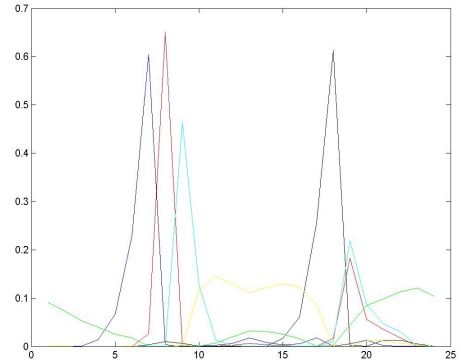
r=3



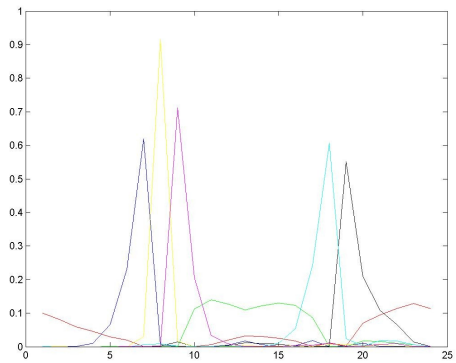
r=4



r=5



r=6



r=7

比较不同的r值分解的结果，可以发现当r=3时语义最清晰：上下班时段、工作时段、其他时段，而增加r时分解结果只是把上下班更细化，直到r=7时分解为：三个上班高峰时段、两个下班高峰时段、工作时段、其他时段（夜晚）。这里我们可以得出结论，车辆的交通基本特征有三种，或者说是车辆出行的目的主要有三种：往返于家与公司、往返于公司间、其他休闲，又可以发现不同的不同车辆上下班的时间会有一定差别。

进一步来解释，这个特征地具体含义，其实一种车辆行驶模式，也就是车辆在一天中不同时间段出现的概率值。分解出来有这么几种不同的类型，这几种类型代表了车辆行驶的主要成分，这些成分的不同组合也就组成了不同具体的车辆的行驶状况。所以这些特征是“基本的”。而并不是说把所有车辆分成了这几种类别。相当于是说，这是他们的主成分。例如当r=6的时候，一辆车的行驶状况就是，“上班早”、“下班晚”、“晚上活动一下”，这三种模式，或许就没有工作时段的“公司业务”，因为不一定所有这个时间都要用车。这就是一种语义解释。

3.3 区域交通情况

3.3.1 方法分析

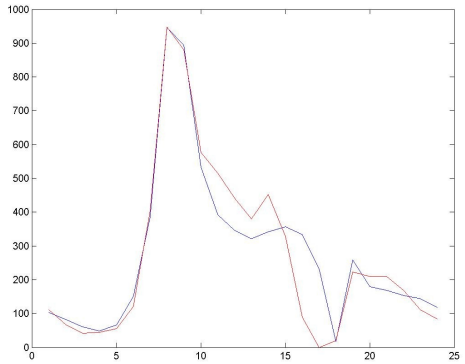
当r=7我们可以看出，车辆的上下班高峰期还是有一定的区别的，所以下一步的任务是分析出不同区域的车流量高峰期差异和主要时间段。首先，要统计出每个时间段内有多少车辆通过某个区域是比较容易的，所以我们可以把区域的交通情况转换成一个向量，就像之前分析的那样（之前用于NMF分解的时候，用的是车辆的行驶情况作为向量），所以也是一个m维向量。要说明的是，这里所谓的区域，指的是某交通拍摄设备那附近的道路区域。具体来说，对于区域i（设备i），其表示的向量为Ei。如果Ei的第j个数值很大，说明在当天的第j个时间段内，交通流量非常大，我们可以称之为高峰时段。所以如果只是观察每个向量哪一个数值最大，就可以粗略反映出该区域的交通流量高峰期时段。这是简洁且行之有效的方法。

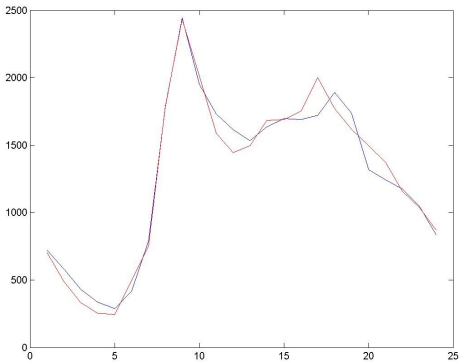
不过，这样的方法是不够全面的，因为不同的车有不同的速度，如果只是看经过的车的次数并不能完全刻画出该区域的流量情况，至少不能完全对等。为了更进一步描述每个区域的交通状况，我们可以利用以上分析出的几种不同的车辆行驶模式。具体说来，假设r=6，那么就有“上班早的”、“上班晚的”、“工作业

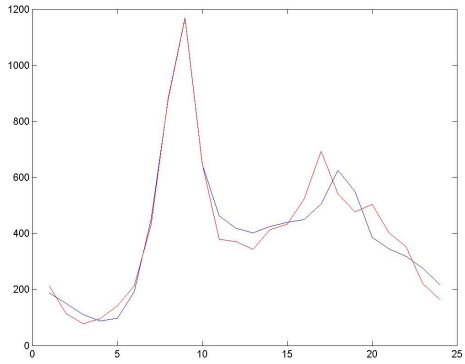
务”、“下班早的”、“下班晚的”、“其他”这几种车辆行驶特征。六种特征分别由6种不同的6个m维向量（m=24）表示，正好区域交通状况也可以由m维向量来表示。相当于是说，既然车辆行驶由这6种基本模式组成，区域的交通状况就是记录的是一天中通过该地方的汽车频率情况，所以车辆的行驶模式的组合也就可以组成区域的交通模式。其实所谓的车辆行驶模式也就是，车辆在一天中不同时间点出现的概率情况。换句话说，既然车辆的主要行驶成分是这几种，他们可以组合成一辆车的行驶情况，那么也就可以组成某个区域的交通情况。例如，某个区域的上班早的车比上班晚的车多一点，公司业务的车几乎没有，晚上娱乐的车又多一点。主要体现出这些成分。又例如，区域A主要反映出上班早的特点，区域B主要反映出上班晚的特点，所以就形成对比说区域A比区域B先达到上班高峰期。这就是所得到的有用的信息。

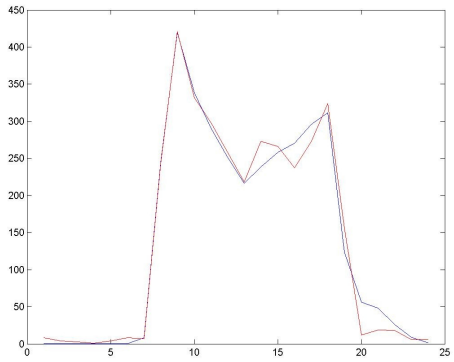
3.3.2 数据拟合

W 表示车辆的基本特征。路口模式也是用一个 24 维向量表示，每一维表示某个时间段 12 天中通过的车的数量。W 矩阵代表了几个基本向量，然后用最小二乘拟合方法，找出这几个基本向量最接近路口向量的线性组合。那么取哪一个 W 是关键，我试着用 r=3 的 W 去拟合路口向量，拟合效果不理想，但是当我用 r=7 的 W 去拟合路口向量，拟合效果很好，这是随机取出的 4 种路口向量拟合情况（红色是路口向量，蓝色是拟合向量）：





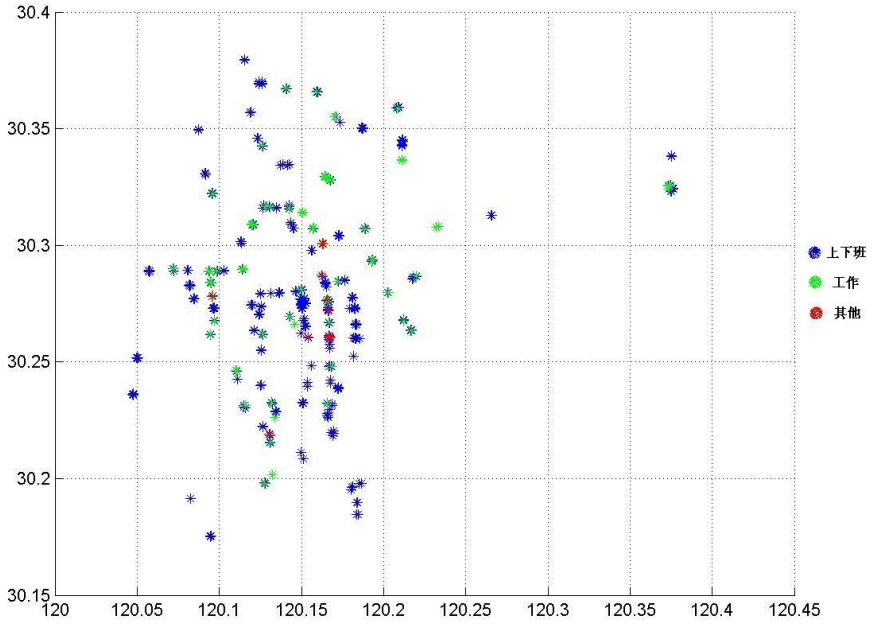




r=7 的拟合效果比 r=3 的拟合效果好，究其原因可以发现 r=3 和 r=7 的 W 的区别就在于后者把上下班的特征分得更细，其实可以解释为，不同路口在不同时间段达到上班和下班的高峰期。

3.3.3 交通路口模式分析

拟合结束后就可以得到不同路口 7 种特征所占比例，因为 7 种特征其实最主要就体现在三种主要特征上面，不管是上班早还是上班晚，不管是上班还是下班，都可以称之为上下班特征，所以合并上下班的五种特征，又变成了三种基本特征。相应的合并比例，就得到了三种特征的比例，从而分析出三种不同的路口（杭州市路口设备经纬度图）：



蓝色表示上下班时段车辆居多的路口，绿色表示工作时段车辆居多的路口，红

色表示其他时段车辆居多的路口。

以下是统计结果：

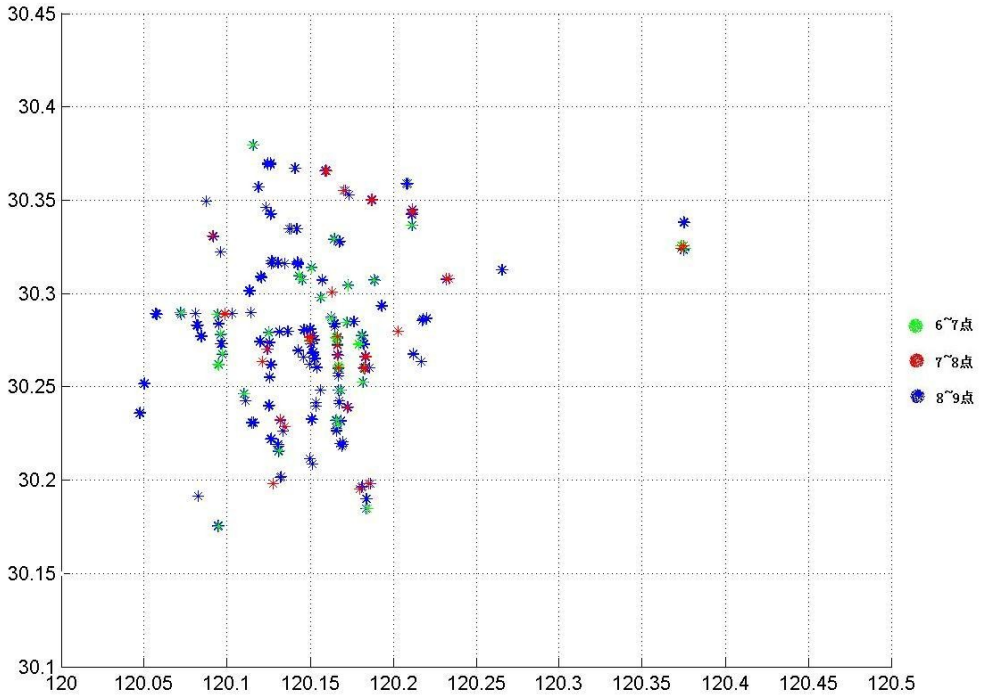
路口总数：756 蓝色路口：659 绿色路口：86 红色路口：11

这表明大多数路口都是在上下班，车辆数达到了高峰，这是符合实际意义的。

3.3.4 高峰时段分析

进一步考察不同路口分别达到上班高峰期和下班高峰期的时间段。也就是比较拟合特征中上班的三种特征比例的大小，以及下班的两种特征比例。

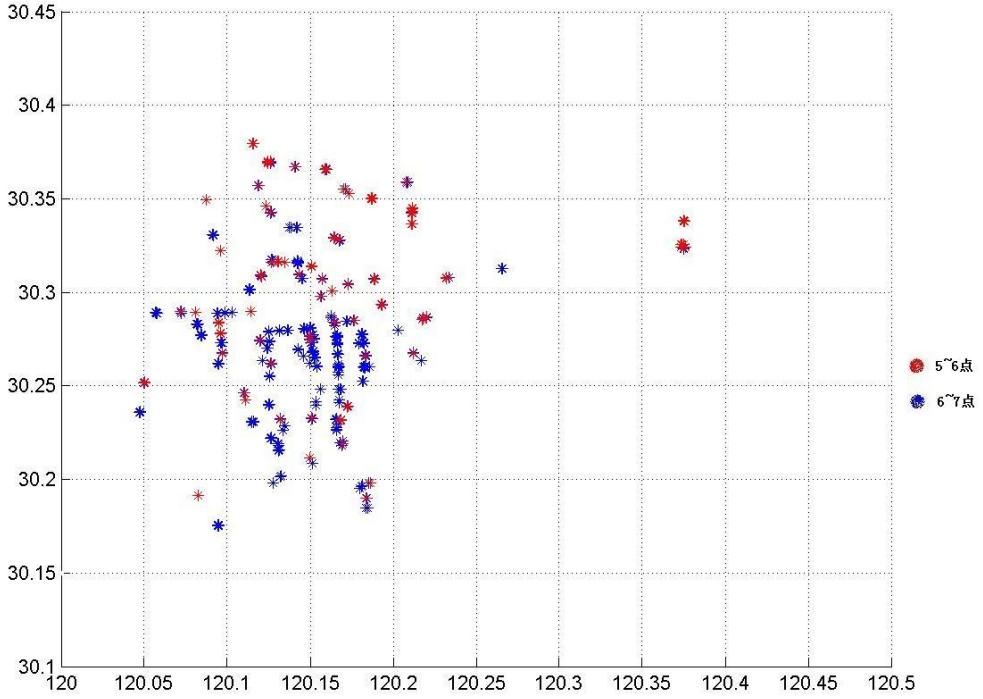
上班高峰期：



以下是统计结果：

6～7 点数量：52　　7～8 点数量：41　　8～9 点数量：663

下班高峰期：



以下是统计结果：

5～6 点数量：138　　6～7 点数量：618

这表明大多数的路口时在 8 点半左右达到上班高峰期，下午 6 点半左右达到下班高峰期，这也是比较合理的。

第四章 结论

直接对路口设备建模进行 NMF 分解得到的结果很不理想，但是对车辆建模进行 NMF 分解却能得到解释性很好的结果，再利用这样的结果去拟合路口设备数据的效果也是比较好的。但是两种不同方法的区别还有待进一步研究。

从实际角度来看上下班高峰时段持续并不是特别长的，但是在 r=7 的分解结果中，有三个不同的上班峰值也就是达到了三个小时，最后分析的路口中 8～9 点达到高峰占了绝大多数，这表明时间粒度太粗，而且最后从经纬度图并不能看出明显的地理特征，所以这一点也有待提高。

参考文献:

1. 高宏娟,潘 晨. 基于非负矩阵分解的人脸识别算法的改进. 计算机技术与发展, 2007(11).
2. C. Peng, X. Jin, K. C. Wong, M. Shi and P. Lio. Collective Human Mobility Pattern from Taxi Trips inUrban Area. PLoS ONE, 2012, 7(4), e34487.
3. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Physics Reports 329: 199—329.
4. Nagel K (1996) Particle hopping models and traffic flow theory. Physical Review E 53: 4655.
5. Esser J, Schreckenberg M (1997) Microscopic simulation of urban traffic based on cellular automata. International Journal of Modern Physics C-Physics and Computer 8: 1025—1036.
6. D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788—791, 1999.
7. Stan Z. Li, X. W. Hou, H. J. Zhang and Q. S. Cheng. Learning Spatially Localized, Parts-Based Representation. IEEE, 2001.
8. 李 乐, 章毓晋. 非负矩阵分解算法综述. 电子学报, 2008(4).

9. 刘维湘,郑南宁,游屈波. 非负矩阵分解及其在模式识别中的应用. 科学通报, 2006(2).

10. Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 1994, 5: 111~126

致谢

感谢中山大学，感谢数学与计算科学学院。学校和学院四年来给我提供的良好的学习环境、便利的科研条件、完善的服务体系，是我得以顺利完成学业的有力保障。在四年的学习生活中，得到了学院各位老师多方面的指导和帮助，感谢他们的大力支持和无私关怀。

最后，衷心感谢为评阅本论文而付出辛勤劳动的各位专家和学者！