“计算机图形学与大规模数据分析”暑期前沿研讨班学习心得

黄一夫

华东师范大学软件学院

10092510437@ecnu.cn

***Abstract*—浙江大学计算机辅助设计与图形学国家重点实验室举办了“计算机图形学与大规模数据分析”暑期前沿研讨班，内容涵盖计算机图形学、数据分析、机器学习、可视分析等。参加研讨班后，发现研讨班的各门课程主要分为理论介绍和应用展望两个部分：理论部分主要借助了最近该领域顶级会议及期刊论文进行阐述，应用部分主要是授课老师的项目展示。该学习心得笔记主要对何晓飞老师教授的流形学习进行总结，再辅以一些其他课程的介绍和感悟。**

***Keywords*—**计算机图形学、数据分析、机器学习、可视分析

I. INTRODUCTION

浙江大学计算机辅助设计与图形学国家重点实验室于2012年7月2日至7月6日在杭州浙江大学紫金港校区蒙民伟楼开设了“计算机图形学与大规模数据分析” 暑期前沿研讨班，内容涵盖计算机图形学、数据分析、机器学习、可视分析等。课程面向相关专业的教师和研究生，既覆盖相关领域基础知识，也涉及大量前沿特色话题，同时将邀请一些国内外专家和企业人士做前沿讲座和学术研讨沙龙。

我参加研讨班后，发现研讨班的各门课程主要分为理论介绍和应用展望两个部分。理论介绍部分主要借助了最近该领域顶级会议及期刊论文进行阐述，点明了研究热点，指出了研究方向；应用部分主要是授课老师的成果展示，成果以实验数据和具体产品进行展现。

研讨班星期一的主题为机器学习，星期二三的主题为计算机图形学，星期四五的主题为可视化与可视分析。在这篇学习心得笔记中，主要针对何晓飞老师教授的流形学习进行回顾，思考与总结，再辅以对一些其他课程的介绍和感悟。

II. METHOD

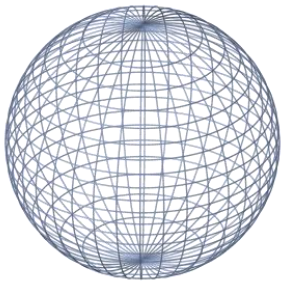
A. 流形学习

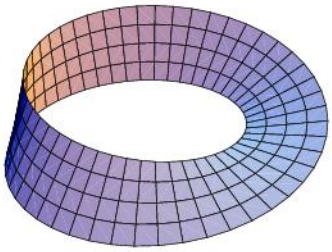
研讨班的第一节课是何晓飞老师教授的流形学习。这门课程并不是对机器学习的基本概念进行介绍，而是站在几何观点上讨论了机器学习。课程难度和深度较大，鉴于我理解有限，之后阐述中不严格的地方敬请谅解。

机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。[1] 我目前所理解的机器学习实质上就是转换函数f的寻找。一般的，有$f: X \rightarrow Y$，但我们平时考虑的X和Y往往是欧氏空间的，然后在欧氏空间上定义该领域的距离表达，再进行计算。这里我们要考虑的是将欧氏空间推广到流形。

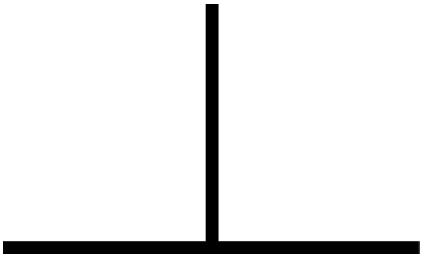
流形，是局部具有欧几里得空间性质的空间。 欧几里得空间就是最简单的流形的实例。地球表面这样的球面则是一个稍微复杂的例子。一般的流形可以通过把许多平直的片折弯并粘连而成。[2] 流形，Manifold = Many + Fold， 代表其为很多曲面片的叠加，要注意的是曲面片叠加而不是拼接，且不为自交。欧氏空间属于流形的特例，任何一个流形都可以嵌入到足够高维度的欧氏空间中(Whitney 嵌入定理)，对于这点我的理解就是：先把流形分段，然后再将其投影到所需的欧氏空间上。

以下是一些流形的例子：





以下是一些非流形的例子：



要在流形上研究问题，还要做一些关于流形的假设，否则可能无从下手。真实数据是这样的：外围欧氏空间的维度很高，数据存在一定的低维内在结构，我们假设数据是位于一个低维子流形上。比如说，把人脸映射到一条曲线上；测地线是弯曲的直线，用来计算流形上两点的最短距离。

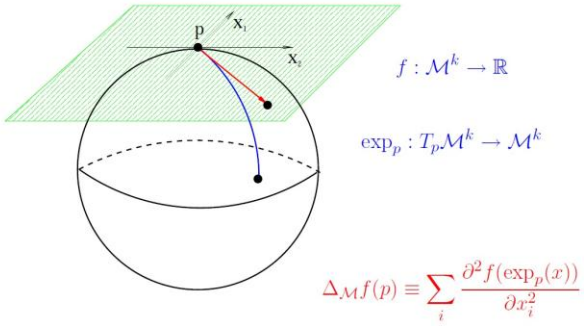
流形有一些特殊性质：不满足平行公设，任意两条测地线(大圆弧)都相交，测地三角形的内角和不一定等于 180 度等。

流形上的学习问题往往是用微分算子表示的微分方程问题，而拉普拉斯算子是微分几何中最重要的微分算子，表示为L或者 Δ，度量了流形上函数的光滑性。

R^3中的拉普拉斯算子：

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

一般流形上的拉普拉斯算子：

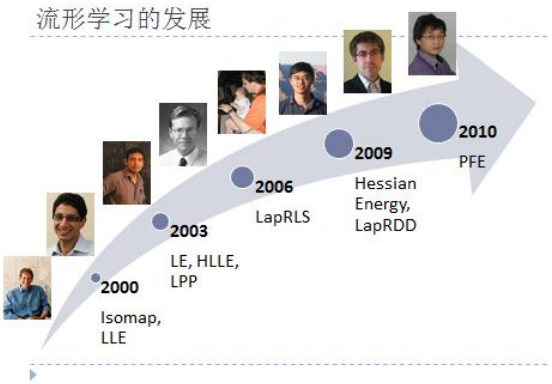


流形模型一般有如下选择：简单模型，对数据要求低，对流形的描述不太准确；复杂模型，对数据要求高，对流形的描述很准确。目前的流形学习基本上都是基于复杂的图模型，但是在研究拓扑结构的时候要用到单纯复形，因为图模型不能刻画高维拓扑。

令数据流形$M \subset R\wedge N$，然后研究在其上的机器学习问题：聚类$f:M \to \{1,\cdots,k\}$，例子有图像分割，社会网络分析，数据挖掘等等，这些都是得到广泛应用的非监督的机器学习；分类/回归$f:M \to \{-1,+1\}$ 或者$f:M \to R$，例子有语音识别，手写体识别，文本分类等等，这些都是得到广泛应用的监督的机器学习；降维:$\boldsymbol{f:M \to R\wedge n, n \ll N}$ ，例子有可视化，应用于后续学习等。鉴于流形学习的重点在于降维，以下讨论的流形学习问题主要针对降维展开。

流形学习起源于 2000 年，其研究热点主要集中降维问题上，多年来有各种各样的降维算法被提出。降维的实现有助于可视化，及简化后续的机器学习问题。

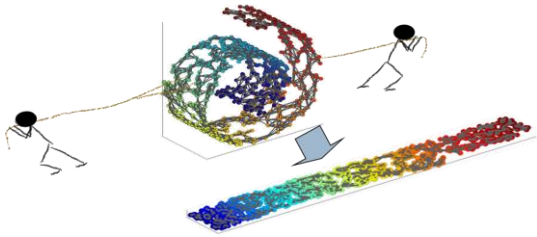
以下是流形学习的概况：



目前，流形学习热点问题是研究数据流形的几何和拓扑。而流形学习的中心问题是降维，即找一个映射从流形到欧氏空间，其经典算法有: ISOMAP, LLE 和 LE。根据流形结构进行学习，半监督学习，主动学习也是比较热门的研究方向。

降维，unfold a manifold，即展开一个流形，同时尽量保持流形的几何结构。

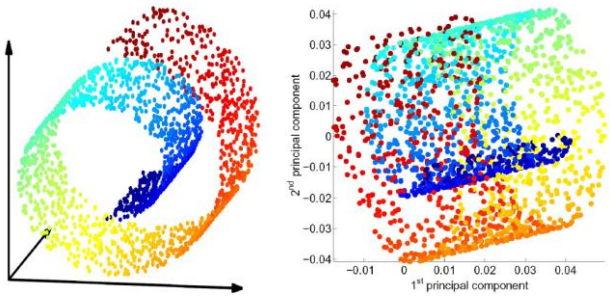
比较形象化的表示为如下图像：



PCA，Principal Component Analysis，传统降维方法。其采用线性投影的方法进行降维，它的目的是使得数据在给定的方向上投影会得到最大的方差，即 $\max_{\|w\|=1} Var\{w^T X\}$，也等价于点 x_n 和投影之后得到的点 $\tilde{x}_n$ 之间的距离最小，即 $\min_{W^T W=I} \sum_{n=1}^{N} \|x_n - WW^T x_n\|^2$。

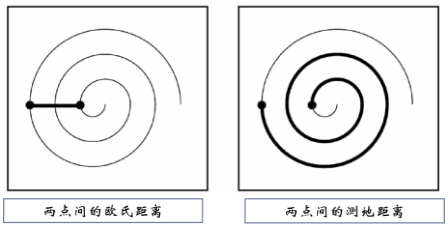
PCA 是到目前为止应用最为广泛的一个降维算法，因为其实现简单且计算量小，在模式识别、金融、生物信息学等各个领域均得到广泛应用。在机器学习本身的众多场景中也通常被用作数据预处理的首要方法，发展出了各种变种（例如 Sparse PCA、Online PCA、Robust PCA、Probabilistic PCA 等）和扩展工作。当流形是一个线性流形时，PCA 得到的结果是最优的；但是 PCA 无法处理非线性流形。

以下是一个例子：



我们可以从例子中明显地看到 PCA 在处理非线性流形的时候丢失了太多几何信息，以至于得到的结果很差。

ISOMAP 则希望在映射过程中保持流形上测地线的距离如下图所示：

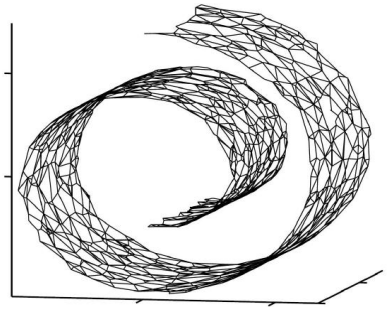


在流形结构未知的情况下，若要根据有限的数据采样来估算流形上的测地线，可以构造邻接图（Graph），用图上的最短距离来近似测地线。两个足够接近的节点之间可以连接上一条边，即

$$W_{ij} = \begin{cases} 1 & \|x_i - x_j\| < \epsilon \\ 0 & otherwise \end{cases}$$

上述式子中的 e 的取值应该按实际数据进行取值，最好能通过机器学习得到较好的参数值而不是人工估计。

下图是对流形数据构造邻接图的可视化：

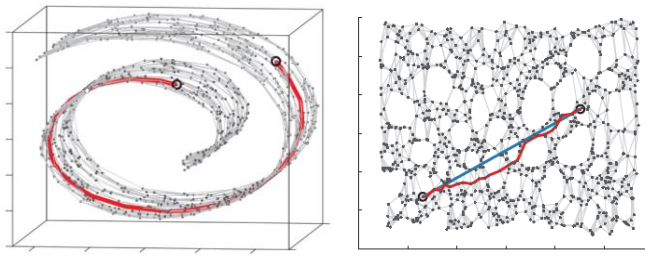


图上两点之间的最短路径（可以用 Dijkstra 或者 Floyd 算法来计算）对应于流形上测地线距离的一个近似值。当数据点趋向于无穷多时，这个估计趋向于真实的测地线距离。这里值得一提的是，这种构造是在数据点大且密集没有空洞的情况下给出的，要是数据在一个局部很稀疏，则会造成邻接图有空洞，或出现分离，这个时候就可能无法计算出最短路径了。

ISOMAP 也可用于降维后的坐标的计算。ISOMAP 先使用 MDS 计算映射后的坐标 y，使得映射坐标下的欧氏距离与原来的测地线距离尽量相等，即

$$\min_y \sum_{i,j} \left(d_{\mathcal{M}}(x_i, x_j) - \|y_i - y_j\|\right)^2$$

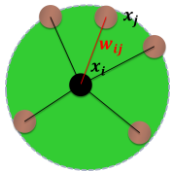
以下是相应的可视化表示：



下面是一个 ISOMAP 的直观示例。将一张图片看成一个数据点，每个像素是一个维度，一张$n \times m$的图像就是一个 nm 维欧氏空间中的一个点。另一方面，数据集中的手的图像只有“张开”和“闭合”以及旋转角度两个自由度，所以这些图片其实分布在一个二维流形上。ISOMAP 在保持流形测地线距离的前提下将数据点映射到二维欧氏空间中，其中在各个位置选取了一些代表性的点将它所对应的原始图片画在旁边。可以看到手的图像在两个自由度上的相似性（距离）得到了保持。

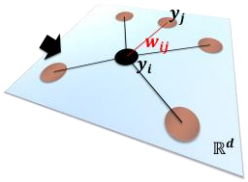
下面介绍另外一个算法 LLE。之前的 ISOMAP 试图通过保持任意两点之间的测地线距离来保持流形的全局几何结构，而 LLE 则从局部来进行分析。“流形在局部可以近似等价于欧氏空间”便是 LLE 分析方法的出发点。

下面是计算最优的重构权重：



$$\arg\min_{w_i} \|x_i - \sum_{j \sim i} w_{ij} x_j\|^2$$

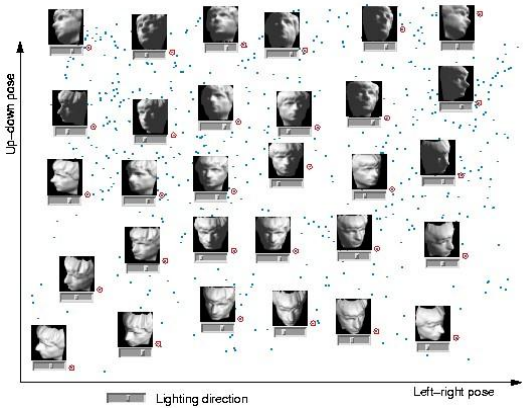
下面是计算最优的映射坐标：



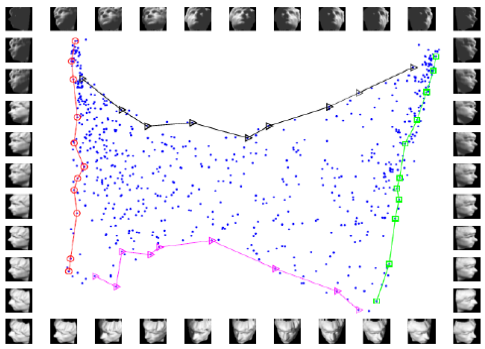
$$\arg\min_{y} \|y_i - \sum_{j \sim i} w_{ij} y_j\|^2$$

将 ISOMAP 和 LLE 进行对比，ISOMAP 和 LLE 从不同的出发点来实现同一个目标，它们都能从某种程度上发现并在映射的过程中保持流形的几何性质。

使用 ISOMAP：

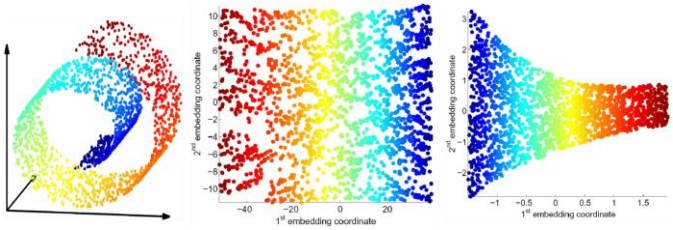


使用 LLE：



ISOMAP 希望保持任意两点之间的测地线距离，LLE 希望保持局部线性关系。从保持几何的角度来看，ISOMAP 保持了更多的信息量。

分别使用 Isomap 和 LLE 对流形进行降维的结果：



从上图可以明显看出 ISOMAP 保持了更多的信息量。

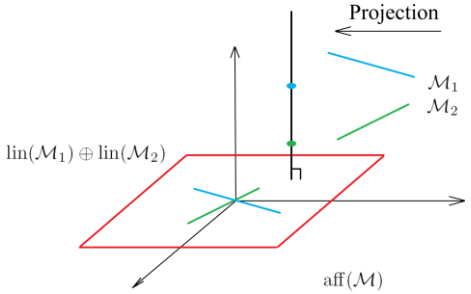
然而 ISOMAP 的全局方法有一个很大的问题就是要考虑任意两点之间的关系，这个数量将随着数据点数量的增多而爆炸性增长，从而使得计算难以负荷。另一方面，随着互联网的发展，我们所面临的数据规模正变得越来越大。比如 Twitter 现在的用户数量已经超过一亿，并且还在飞速增长。诸如此类的巨型结构使用全局方法进行分析正在变得越来越不切实际。因此，以 LLE 为开端的局部分析方法的变种和相关的理论基础研究逐渐受到更多的关注。

还有一个叫 LE 的算法，希望保持流形的近邻关系。将原始空间中相近的点映射成目标空间中相近的点目标函数：E(y)=w_ij (y_i-y_j)^2，约束条件：y^T y=1，去除任意的缩放，转化成一个特征向量问题：$Ly=\lambda y$。LE 的求特征向量问题对应于连续的时候求拉普拉斯特征函数问题：先构建近邻图；然后计算每条边的权重(不相连的边权重为 0)，其中热核权重：w_ij=exp(‖ x_i-x_j ‖ ^2/σ^2)，0-1 权重：w_ij=1；最后求解特征向量方程，$Ly=\lambda y$，将点x_i映射到(y_1 (i),⋯,y_d (i))

对 LE 算法分析可知，LE 希望保持数据的近邻关系，实际上并没有最大限度的去保持数据的几何特征。LE 的准则非常适合于做聚类。

Locality Preserving Projections (LPP)，跟 LE 一样的准则，但是限制函数是外围欧氏空间的线性函数：$f(x)=\boldsymbol{a}^T x$,$x\in R^N$，最终转化成求解如下特征向量问题：$XLX^T \boldsymbol{a}=\lambda XDX^T \boldsymbol{a}$，其中$X$=($x\_1$ ⋯x_n)是数据矩阵。

LPP 可以将位于平行仿射凸包的流形分开(Binbin Lin, 2010)：



对比 LPP 和 PCA，可发现 PCA 考虑的是全局统计信息，LPP 是第一个考虑流形结构的线性方法。何晓飞老师在上课过程中还举了学多人脸识别相关的例子，这里限于篇幅，就不做展开了。

最后谈谈流形学习的展望和挑战。研究数据的几何和拓扑，对于人们认识数据和处理数据具有本质意义。现有方法对于数据的要求比较高，对噪音的情况处理能力不够。特殊的向量场反应流形的几何和拓扑，同时跟流形上的函数密切相关。证明向量场方法的优势，有很多基础的理论问题需要解决。结合流形结构的学习问题，需要在流形上发展相应的统计概念以及学习理论。互联网时代，大规模数据的流形学习，流形学习算法往往是要求一个整体的矩阵分解，很难处理大规模数据。

B. 其他

以下主要对一些其他课程进行笔记型的简述。

跨媒体理解中的结构性学习

首先介绍图像检索（从元数据搜索到以图搜图再到图像理解的三个发展阶段），再介绍了跨媒体的检索与理解，然后介绍了 Marr 理论与 Gestar 理论，最后推出结构性机器学习，并分析了其与传统型机器学习的区别，而且给出了已较广泛使用的模型方法如 structural svm 等。

矩阵分解及其应用

矩阵分解的实质就是将一个大矩阵分解成两个小矩阵的乘积，在此之上可能有很多的变形以针对不同的应用。然后介绍并展示了矩阵分解在图像恢复，推荐系统，信息检索等方面的应用。之前在流形学习部分也提到，在互联网时代，大规模数据的流形学习，流形学习算法往往是要求一个整体的矩阵分解。

大数据的智能处理

这是一个 Introduction 型的课程，主要从各种应用对大数据的职能处理进行介绍，比如 PAROS 系统。

移动轨迹数据分析与挖掘

首先介绍了一些应用：连续轨迹还原，其基于 ST 匹配算法；道路重建，绘制；异常轨迹检测，比如说出租车绕路，基于字符串匹配；语义地点提取；本征行为分析。然后分析了一篇 Science 上的文章，主要含以下几个方面：1.移动步长规律，服从幂律分布；2.访问点规律，探索性，回溯性；3.移动可预测性，移动轨迹熵。由此可见 Science 上的文章倾向于去发现规律的发现，而不是针对各种应用在领域上的探索。最后讲述了社会活动模式挖掘 ：1.移动与社会关系；2.移动与社会事件；3.移动与区域功能。其可以应用于监控道路环境，改进交通服务，分析城市规划等。

基于视频的三维建模

主要介绍了如何通过一段视频，来重构三维模型，实质上是从多个视角的图像来给予三维模型中的各个像素点以深度。介绍了如下概念及算法：Feature tracking SIFT, KLT； Graph-cut； belief propagation； PMVS； Poisson suffice recognition。

隐函数曲面造型和动画技术

自由曲面分为：参数曲面，多边形曲面，细分曲面和隐式曲面。其中隐式曲面的关键问题是，表示，计算和运动控制。在实际应用中最小单位一般都使用 metaball（元球造型）表示。之后还介绍了：能量函数；法向函数；收缩函数；复合函数；复杂骨架卷积等。

数据可视化基础

可视化的功能一般可理解为：记录信息，支持对信息的推理和分析，和信息传播。目前的数据科学主要针对的是大数据，其有我们常说的 3V 特性，并且一些优秀的应用也层出不穷：用于国土安全部的 DHS，CVADA 可视化数据分析，Insight 情报分析，处理半结构非结构数据的 XDATA 等。

Data 通过 ETL（提取，转化，加载）到 data warehouse，再到数据产品，BI，和分析。这里有一点想提的是：目前可视清洗是一个比较热门的话题，脏数据的不完整，冗余等信息让人头疼，数据清洗本身是一件工作量极大的事情，因为事先不知道数据的分布，需要慢慢摸索尝试，而可视化可以缩短这段耗时。可视化的基本表现形式有：统计图表，其中数据轴范围的取舍很重要；尺度；百分比，比例数据；stream graph，stack graph；散点图矩阵；盒须图等等。

高维数据的可视化基本手段：数据表；平行坐标，聚类，透明度，用在十几维上时不错；星型散点图；切尔诺夫脸谱图；马赛克图（最多应用于五六维）；平行集；属性直方图（用颜色，纹理等描述）；散点图，紧凑的像素显示（用正方形编码，按相似性进行排列）。

降维，做交互分类，聚类。低维嵌入：降维算法在之前的流形中提到过，线性方法有 PCA, MDS, NMF；非线性方法有 LLE, ISOMAP, Charting。目前其中 MDS 应用较为广泛：输出与输入距离尽量一致。

树的基本表示手段：有根树，基于缩进；节点链图；Enclosure；双曲空间；树图；分层显示。网络的基本表示手段：sugiyama 类显示，原生顺序的树；边聚类，深浅；正交图；环形排列；嵌套排列等。另外可视化的实践方面最好从 tool 入手，比如说 graphviz 和 processing。

交互的准则及一般步骤：1.select 标记感兴趣，跟踪其变化；2.explore 显示不同东西；3.reconfigure 显示不同排列；4.encode 显示一个不同的表示方法；5.abstract/elaborate 显示更多更少细节 details-on-demand；6.filter 显示符合条件的某些东西；7.connect 显示相关的项目；8.overview。

商业与社交数据可视化

商业数据方面主要介绍了淘宝的可视化。淘宝的数据特征：海量，复杂，高价值。目前比较热门的应用有 CatMap，CatLink，TaoTrends，Travel Trends，淘小微，Head Tracking 3D，DataV.Js 可视化组件库，shu.taobao.com，datavlab.org 分享交流平台等。

社交网络一般是基于节点链接图或邻接矩阵来完成的，这里主要描述的是在复杂的社交网络中做 Community Detection，同时加入 context information，采用 Modularity 优化方程进行聚类。

实战云计算

从实战入手，讲述了用 1500 台阿里云的服务器进行渲染工作的经历（Render.aliyun.com），点明了云计算的四大关键：MapReduce，VM，云存储，DB。也将其粗略分层：SaaS，PaaS（X App Engine），IaaS（Google compute engine）。

可视分析

先着重介绍了可视化的可视分析的区别，在这方面我的理解是：可视化就是将数据用合适的可视组件展示出来；可视分析是将人的分析加入可视化的过程中，进行迭代获得知识。然后就是以下经典实例的介绍：Bullseye，Flickr tag cloud，wordle，themeriver，document card，word tree，phrase net，galaxiex，jigsaw，facetatlas，palantiv（open source），infovis，vast。

信息可视化

这节课是北京大学可视分析小组的袁晓如研究员教授的。主要针对其成果 PKU weibo visual，进行了实时展示，展现出来信息可视化。对比了 Info vis 和 science vis。同时也指出了 High dimension 的可视化方法，比如 MDS,PCP 等。最后介绍了 WYSIWYG 理念。

科学可视化

介绍了 VTK：visulization toolkit。SCIrun：problem。着重点在于数据的表示，数据结构。

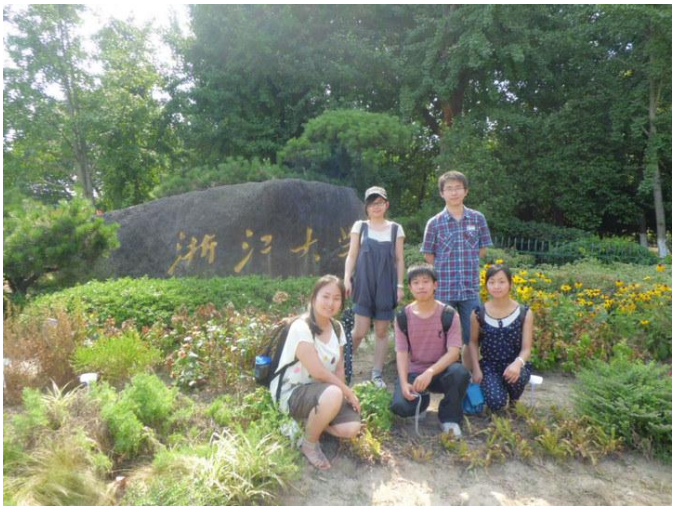
III. CONCLUSIONS

参加了浙江大学“计算机图形学与大规模数据分析”暑期前沿研讨班后，我在知识面和深度方面得到了一定的提升。这次安排的课程中，我对机器学习，数据挖掘，可视化等方面课程很感兴趣，下课后更是和老师进行进一步交流，感觉收获颇丰。机器学习方面主要是对讨论问题的空间进行了扩展，数据挖掘方面介绍了许多有新意的应用，可视化方面主要是可视基础的夯实。

ACKNOWLEDGMENT

本次研讨班是盛斌老师推荐参加，且车费及住宿费来源于盛斌老师开放课题的资助，在此对盛斌老师进行感谢。除此之外，还要感谢张双力，季雨航，李真真，甘振业的旅程安排及同行。

最后附合影一张：



REFERENCES

[1] http://zh.wikipedia.org/wiki/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0

[2] http://zh.wikipedia.org/wiki/%E6%B5%81%E5%BD%A