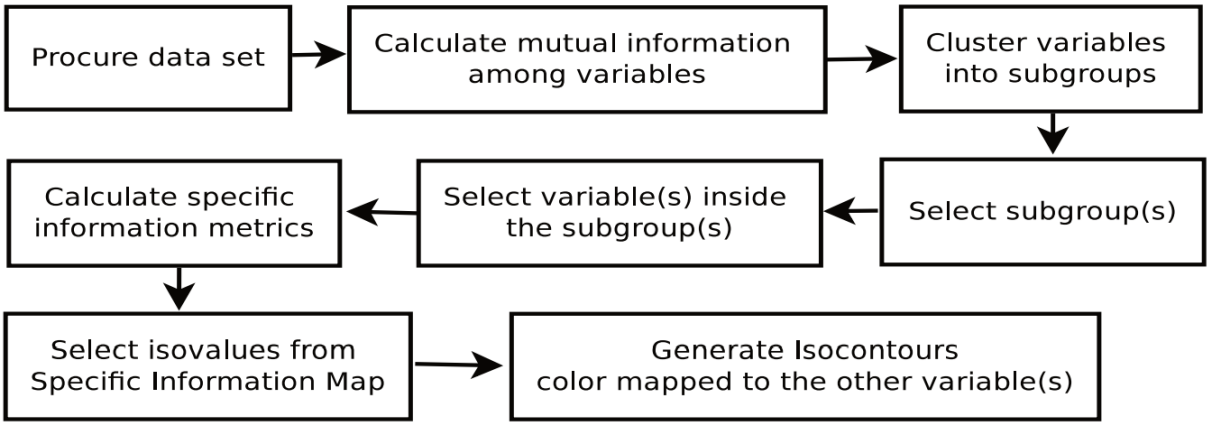
PKU Visualization Blog 北京大学可视化与可视分析博客

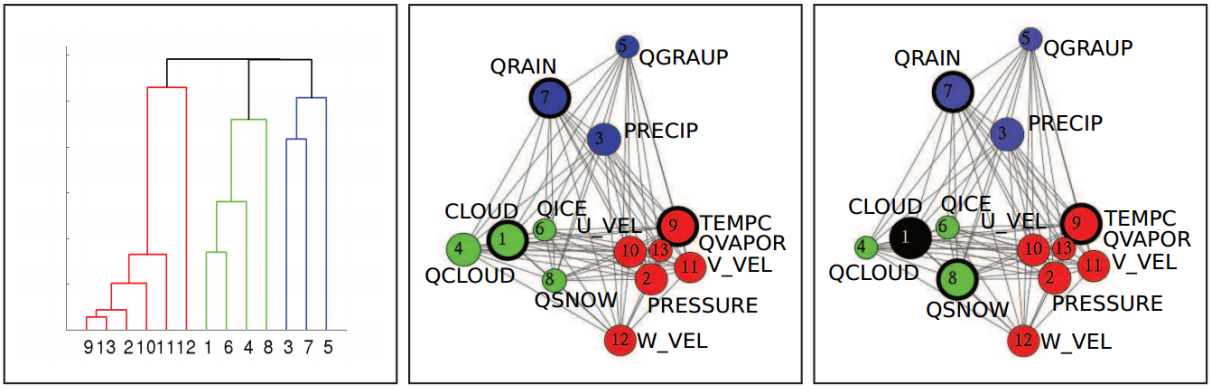
一种探索多变量数据集的信息感知框架(An Information-Aware Framework for Exploring Multivariate Data sets)

作者: Jiang Zhang 日期: 2013年11月16日

多变量数据集的探索是科学可视化中的一个重要研究方面，可以让人们对多个变量之间的关系进行深入地了解。在单变量系统中，isocontour是一个很常用的手段，可以用来揭示相同标量值的区域。而在多变量数据集中，由于多个变量之间的相互依赖型，等高线可以展示与其关联的变量的信息和交互方式，因此也成为了多变量数据研究的一个重要内容。在这篇文章[1]中，作者提出了一种信息感知的框架来引导用户进行多变量数据的探索。文章使用了信息论的方法来计算变量之间的互信息，还使用了特定信息(specific information)来计算某个变量的标量值与其他变量的相关性，从而使用isocontour对其不确定信息进行探究。除此之外，文中还使用了直观的交互界面，如平行坐标、散点图等。

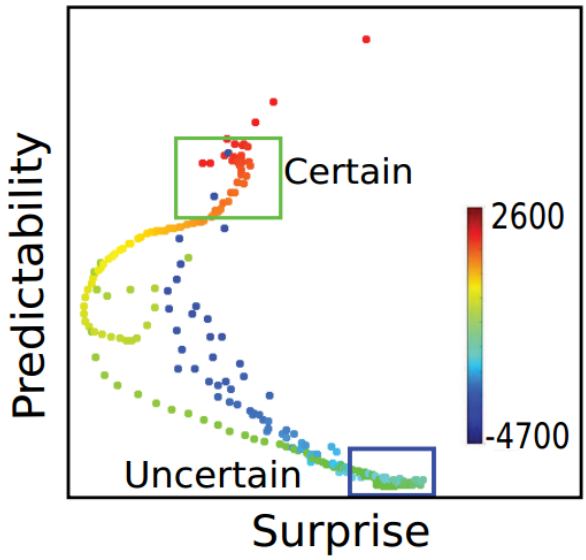


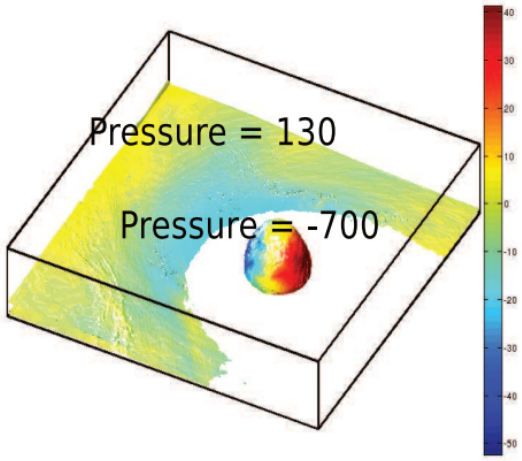
系统的流程如上图所示。首先可以计算出变量两两之间的互信息，根据这些互信息，可以将变量细分成更小的子组。之后，使用联合熵(joint entropy)选出有代表性的子组，在该子组中，使用条件熵继续筛选出含有最大信息量的变量，可以以此作为引用变量(reference variable)，用来计算出特定信息度量(specific information metrics)，然后探究引用变量的标量值对其他变量不确定性的影响。接下来看看具体操作。



基于变量之间的信息重叠，两个变量之间的互信息可以用它们的概率分布计算出来（公式1）。然后对于系统中所有的变量，作者使用了一个图布局来描述它们之间的关系。结点表示变量，无向边表示结点间的互信息。结点间的引力与互信息成正相关，斥力与互信息的倒数正相关。文中使用了一个层次聚类的方法，初始时每个变量都是一个聚类。如果两个聚类之间具有最大的相似度，则将它们组成一个新的聚类。如上图所示，这里面一共组成了三个聚类，分别使用不同的颜色表示。用户选择一个变量后，其他变量的相对重要性会实时更新并反映在图中，以便用户可以进行进一步地探索。

在将变量进行聚类后，每一个组都可以计算出它的联合熵，据此选出联合熵最大的组。在该组中，使用条件熵计算出单个变量对该组总的不确定性的贡献。条件熵高的变量可以为引用变量提供选择。之后，就可以计算特定信息度量了。特定信息是用来度量引用变量的(X)一个特定标量值(x)对另一个变量(Y)的不确定性的影响。特定信息有两种度量方式，一种叫做Surprise（公式2），高的Surprise值表示一些罕有出现的y值的出现概率会因为x的观测而增大。另一种叫做Predictability（公式3），表示观测到X的一个标量值x后Y的不确定性的减少量。在本文中，两种度量方式是结合起来使用的。





除了使用图布局来展示变量之间的关系，作者还使用了两种探索方式。一种是在数据域的探索，利用平行坐标来表示变量关系，使用互信息来组织竖直轴的顺序。为了得到最大的信息量，文章还在子图中使用了最小权值汉密尔顿路径来使得边权值之和最小，因为边权值是与互信息的倒数正相关的。为了较少视觉混乱，刷选技术也被用来对感兴趣区域的选择。文章引入了两个额外的度量，一个是Iuncert（公式4），用来辨别高的Surprise值和低的Predictability值，另一个是Icert（公式5），用来辨别高的Surprise值和高的Predictability值。除此之外，另外一种探索方式是在空间域的探索，利用散点图将引用变量的标量映射到Surprise-Predictability空间。用户可以在散点图中选择一个点或者区域，相关的结果会在一个isosurface视图里面呈现。如上图所示，散点图中蓝色区域对应于压力值为-700，它有比较高的变化性。而对于压力值为130（绿色区域），其变化性相对少很多。

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \quad (1)$$

$$I_1(x;Y) = \sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{p(y)}. \quad (2)$$

$$I_2(x;Y) = H(Y) - H(Y|x) = -\sum_{y\in Y} p(y)\log p(y) + \sum_{y\in Y} p(y|x)\log p(y|x) \quad (3)$$

$$I_{uncert} = I_1/I_2. \quad (4)$$

$$I_{cert} = I_1 * I_2. \quad (5)$$

总的来说，这篇文章使用信息论的方法计算变量间的互信息并据此来计算两个特定信息度量，从而探索引用变量的特定标量值与其他变量的关系。除此之外，文章还提供了图布局以及平行坐标和散点图来展示变量关系。不过，对于文中的一些做法还有一些疑惑，比如平行坐标的使用是为了突出什么，散点图和isosurface视图如何通过引用变量的特定标量值进行关联，等等。所以，也欢迎和对这篇文章同样感兴趣的同学一起进行讨论。

参考文献

[1] Ayan Biswas, Soumya Dutta, Han-Wei Shen, Jonathan Woodring. "An Information-Aware Framework for Exploring Multivariate Data Sets". Visualization and Computer Graphics, IEEE Transactions on, vol.19, no.12, pp.2683,2692, Dec. 2013

论文报告 information theory, isosurface, multivariate uncertainty

← 基于贝叶斯模型平均的集合数值模拟预测不确定性的表征和可视化(Characterizing and Visualizing Predictive Uncertainty in Numerical Ensembles Through Bayesian Model Averaging)

通过使用离散小波变换的积分直方图来进行高效的局部统计分析 (Efficient Local Statistical Analysis via Integral Histograms with Discrete Wavelet Transform) →

评论关闭。

RSS 订阅

功能

登录
文章RSS
评论RSS
WordPress.org

链接

北京大学可视化与可视分析研究小组主页 – PKU Vis Home Page
北京大学可视化研究维基 – PKU Vis WIKI

分类目录

应用
新闻
未分类
活动
研究
论文报告

标签

ChinaVis graph interaction PacificVis pacificvis2019 pvis2016 不确定性 主题模型 交互 交互设计 人机交互 会议 体可视化 体绘