ノンパラメトリックな相対リスク推定手法を用いた空間データマイニング

大阪市立大学大学院経営学研究科　付属先端研究教育センター　井上 暁光

近年、多様な空間データが入手可能となり、それらのデータから有用な知識やパターンを発見し意思決定に応用する機会が増加している。空間上の点の座標と共に観測される各点の種々の属性は、空間上の事象の発生頻度のパターンを発見するための有用な情報となりうる。本研究の目的は、座標と複数の離散的な属性値が空間上の各点に与えられている時、それらの属性によって層別化した空間上の点の分布の相違を計算し、属性と対応する分布の相違を重要性に応じて可視化する方法を提案することである。本研究では、空間上の相対リスク推定に用いられる、カーネル密度推定法と密度比推定法を用いた新たな手法を提案する。

2 次元ユークリッド空間上で定義される n 個の点を表す $n \times 2$ の行列 $\mathbf{X}$ と、n 個の各点に定義される離散的な属性を 0 または 1 の値をとる m 個のダミー変数として表した $n \times m$ の行列 $\mathbf{Y}$ を考える。j 番目のダミー変数に対応する相対リスク関数の推定値は、

$$\hat{r}_j(\mathbf{x}) = \log \frac{\hat{f}_j(\boldsymbol{x})}{\hat{g}_j(\boldsymbol{x})}$$

で表される。ただし、$\hat{f}_j$, $\hat{g}_j$ はそれぞれカーネル密度推定量であり、それぞれ、

$$\hat{f}_j(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} \boldsymbol{Y}_{i,j}\, K_{h_j}(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_{i=1}^{n} \boldsymbol{Y}_{i,j}},$$

$$\hat{g}_j(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} (\boldsymbol{Y}_{i,j} - 1) K_{h_j}(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_{i=1}^{n} (\boldsymbol{Y}_{i,j} - 1)}$$

として定義される。ここで K_{h_j} はカーネル関数である。推定された密度関数および相対リスク関数から、それぞれのダミー変数の重要度を $\hat{f}_j(\boldsymbol{x})$, $\hat{g}_j(\boldsymbol{x})$ の異質性を表現する関数 $I_j(\hat{f}, \hat{g})$ によって計算する。また、ダミー変数間の類似性を、2 つの相対リスク関数の類似性を表現する関数 $S_{j,k}(\hat{r}_j, \hat{r}_k) = \left| \int sgn(\hat{r}_j(\boldsymbol{x})\ \hat{r}_k(\boldsymbol{x})\ ) \boldsymbol{dx} \right|$ によって計算する。計算された $I_j, S_{j,k}$ を利用し、ダミー変数の重要性とダミー変数間の類似性を可視化する。詳細および適用例は当日示す。

参考文献

Davies, T.M. and Hazelton, M.L. (2010), Adaptive kernel estimation of spatial relative risk, Statistics in Medicine, 29(23) 2423-2437.