生物統計学/疫学・予防保健学抄読会

大規模データに対する樹形モデルの方法論

2016/02/10　博士課程 1 年　奥井佑

一般的に認知症や生活習慣病などの多因子疾患は遺伝的要因や社会科学的要因、生活習慣などが関連しあって生じる。そのため、疾患に対する予測・因果モデルや、疾患と関連する要因間のパス図を作成する際には、分子生物学的データから生化学データ、社会環境的要因までを含めた多角的な情報を含んだ大規模データを解析する必要性がある。

この場合の大規模データは異種のデータが混在した高次元データであり、扱う際の統計学的課題として変数選択が問題となり、データの多くを占めるノイズを除去して重要な変数を抽出する必要がある。その際、各個体について全ての変数が観測されていることは稀であるため、変数の欠測を適切に補完した特徴抽出が求められる。また、異種類のデータを無秩序に混ぜるのではなく、疾患の発症経路など事前に分かっている医学的見地をもとに解析することが望ましい。

このような大規模データを処理する方法として機械学習による手法を用いたデータマイニングが行われる。機械学習のうち、CART(classification and regression trees)は予測変数空間を再帰的に 2 分割して予測モデルを作成する方法であり、解釈が容易である点や、外れ値や変数の単調変換に頑健であるといった点、異種類の変数が混在するデータの扱いに優れており変数選択効果もあるといった多くの長所を持つ。一方で、CART は定数近似に基づく方法であり予測精度が低いことから、付随して複数の学習モデルを併合する集団学習が用いられる。集団学習により CART の予測精度を上げることができ、機械学習手法の中でも有用な学習方法であると考えられる。

本抄読会では、CART と集団学習の各方法について簡単に説明し、異種データが混在する高次元データへの応用可能性について考える。

参考文献

1. Friedman JH, HastieT, Tibshirani R.Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics* 2000; 28: 337-407.
2. Breiman L. Random forests. *Machine Learning* 2001; 45:5-32.
3. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning; Data Mining, Inference and Prediction 2ed*. Springer: New York, 2009.