

欠測補完のためのドメイン設定と k-近傍法について

独立行政法人 統計センター 坂下 佳一郎

目的

一般に統計調査から得られる観測データには欠測が起こることがあり、そのまま集計すれば結果に偏りが生じる場合には補完 (imputation) 処理を行うことになる。補完の手法には観測項目により設定されるドメイン (補完実施区分) 毎に補完処理を施すものがある。例えば、政府の統計調査の集計実務において欠測補完を実施する際には、前回調査の結果等から得られる経験を基に、産業分類により設定されたドメインごとになされる場合がある。

しかし、新規調査でドメイン候補が多い場合、ドメインを設定する手法ではその設定に多大なコストを要する場合がある。そこで、本研究では、k-近傍法を適用することでカテゴリ項目のほか数値項目の欠測補完をドメイン設定なしに行うことを試みる。その上で k-近傍法を適用した場合とドメイン設定をした上で欠測補完を実施した場合との関係について比較する。

数値項目の欠測補完に対する k-近傍法の活用

通常、k-近傍法はクラス分けなどの質的項目の補完に使われることが多い。その場合は補完対象データの近傍 k 個の当該質的項目の値によるいわゆる「多数決」で補完値を決定することになる。

一方、本研究では例えば、企業の売上額など、数値項目の欠測に対して k-近傍法で補完を試みる。この際、近傍 k 個の当該数値項目の値を 1 つの補完値にどのように合成するかを表す合成関数の検討が必要である。本研究ではこの合成関数として「近傍 k 個の中央値」、「近傍 k 個の平均値」、「距離の逆数で重みづけした近傍 k 個の加重平均値」を中心に採用する。

本研究における k-近傍法の試算環境

本研究における k-近傍法の適用では統計ソフト R の VIM パッケージの関数 `kNN_work` を必要に応じて調整して活用する。また、数値項目の欠測補完に k-近傍法を適用する際に、先に述べた合成関数のほかに事前に検討すべき点が 2 点ある。 k の値及びデータ間の類似度を表現する距離関数である。 k については 10 以下の値を中心に複数の値を使用する。距離関数については関数 `kNN_work` で採用されている Gower 距離を用いた。計算結果の詳細については、当日、報告する。

参考文献

- Gower, J. C. (1971) A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, Vol. 27, No. 4, pp. 857-871.
- Kowarik, A. and Templ, M. (2016) Imputation with the R Package VIM. *Journal of Statistical Software*, Volume 74, Issue 7