ビッグデータ対応の二クラス分類KY(K-step Yard sampling)法の開発と展開

株式会社　インシリコデータ　湯田　浩太郎

１．はじめに

サンプル群を二クラスに分類する手法は、データ解析において重回帰手法と並び極めて重要な解析手法である。サンプル数が少ない時、完全分類の実現は可能であるが、サンプル数が増大するに従って完全分類の実現は不可能となる。現在は、データベース等の発展に伴い、データ解析で扱うサンプル数は急速に増大している。このようなサンプル数増大に対応し、かつ強力な解析能力を有する多変量解析/パターン認識手法を開発したので報告する。

２．データ解析実施基本概念：KY(K-step Yard sampling)法

現在の殆どの多変量解析/パターン認識手法は一度のみの計算で結論を出す。このため、サンプル数が増えるにつれ、完全分類を実現することは極めて困難となる。多段階で行う手法としてバイナリーツリー等があるが、サンプル群の分類は最終のリーフに至るまで決まらない。

KY法は、以下の二つの特徴を有するアルゴリズム（手順）の名称である。

１．サンプル群を、正分類サンプル群と分類が決められないグレーサンプル群に分ける。

２.グレーサンプルと判定されたサンプル群を初期サンプル群として、再び１の操作を繰り返し、正分類サンプル群とグレーサンプル群を分ける。

この二つの操作を一段階（セット）とし、これを繰り返し実施する。最終的にグレーサンプル群が０となった時点で解析が終了する。これにより、全サンプル群が完全分類される。このKY法で、正サンプル群とグレーサンプル群への分類に使われる多変量解析/パターン認識手法は、従来から展開されている二クラス分類手法が適用される。従って、KY法は従来手法の適用手順を変えただけの“メタ解析手法”である。この故、新規の二クラス分類手法が展開されればそれを用いてKY法として展開可能である。また、各段階で異なる二クラス分類手法を混在させても問題ない。

３．結論

KY法はリサンプリングと段階的な繰り返し操作を基本とする手法の総称となる。現在、二クラス分類KY法として３種類[1,2,3]開発されている。特に、「モデルフリー二クラス分類KY法[3]」はアルゴリズムが単純なため、プログラム開発が容易で、予測用の判別関数を予め用意する必要がなくメンテナンス性にも優れている。また、KY法を適用した重回帰手法の開発[4]も行われ、良好な結果を得ている。今後はKY法をクラスタリング等に適用することも検討中である。

KY法の基本アルゴリズムより、理論上サンプル数がビッグデータレベルまで増えても二クラス分類KY法では常に完全分類を実現できる。また、重回帰KY法では極めて高い相関係数を実現することが可能である。詳細は発表の場において説明／討論する。

４．参考文献

1. United States Patent　7,725,413 Yuta　May 25, 2010

2. United States Patent Application 20100241598　Kind Code A1 YUTA September 23, 2010

3. United States Patent Application 20110137841 Kind Code A1 YUTA June 9, 2011

4. United States Patent Application 20110208495 Kind Code A1 YUTA August 25, 2011