多様性に富んだ多量データを S言語を用いて読み解く

ユーザー訪問／群馬大学工学部情報工学科

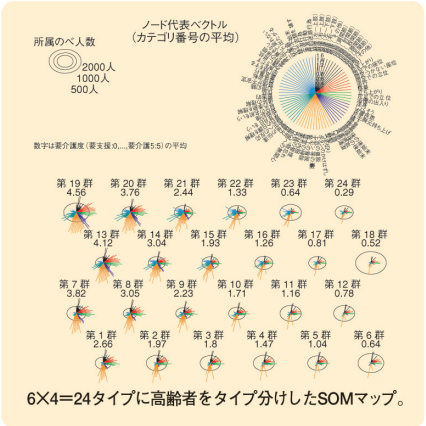
情報ネットワークの進化に伴い，デジタル化されたデータが大量に収集／蓄積されるようになった。多様性を含むこうした多量のデータを，均質なグループになるようにデータを分別するための方法を層別するために，群馬大学工学部の関庸一教授はS言語を用いている。

群馬大学工学部情報工学科
教授　工学博士　関庸一氏

多量データから知識を引き出す データマイニングの方法の開発

群馬大学工学部情報工学科教授・関庸一氏は，統計データ解析モデルの推測方式や，確率モデルの理論解析とシミュレーションによる評価／分析などの研究を行なっている。

「ネットワーク技術，センサー技術の発展とともに，多量の情報を広範囲から即時に双方向で収集できるようになりました。そのため，いろいろなところでデジタル化されたデータが，多量に収集／蓄積されるようになってきています。たとえばスーパーやデパートなどのレジでは，バーコードで商品を調べて会計をします。そこでポイントカードやクレジットカードを使うと，カードの持ち主が特定され，買い物の内容と付き合わされて，いつ，どの店舗で，どのお客が，どんな商品を購入したかというデータが収集されることになります。この情報は，ネットワークを通じて集められ，在庫の管理や，売れ筋商品の把握だけでなく，個々の顧客の好みを知るためなどに使われることになります」



6×4=24タイプに高齢者をタイプ分けしたSOMマップ。

現在は，このようにデジタル化された大量な情報を収集・蓄積・加工し，そこから有益な情報を引き出せる環境が整っていると関氏は語る。しかし，こうして集められたデータの1件1件は，そのときの個人の気まぐれの結果であると同時に，少ない情報しか持っていない。その上，広い範囲から情報が収集されるため，さまざまな個人類型が混在した多様性に富んだものになる。また，想定される多様性には，単に各時点，各個人の多様性だけでなく，時間軸上での変化タイプの多様性もあることから，従来の単一の統計モデルではその多様性に対応できないという問題が生じる。

「このような多様な個人を適切に分別・層別することができて初めて，従来の精緻な統計モデルを適用することが可能になると考えています。そこで，古典的な統計的方法論が適用できるような均質なグループにするための手法を採用して，実データに対して有効な方法論の研究を進めています」

関氏の研究室では，多量データから知識を引き出すデータマイニングの方法の開発と活用に取り組み，多量データの構造を理解する方法を見出そうと研究を行っている。

均質なグループになるよう SOMを用いてデータを分別

「物事をわきまえていることを分別があるといいますが，多様性を含む多量のデータで最も基本となるのは，それぞれができるだけ均質なグループになるよう，データを分別することです。そのひとつの方法にSOM（Self-Organizing Maps：自己組織化マップ）といわれる方法があります。この方法は，人間の神経細胞に倣って開発されたニューラルネットのひとつで，沢山のものから代表的な類型を構成して二次元平面に並べてくれます。たとえば，複数店舗の利用に関する生活習慣に関心があれば、各顧客の各店舗への来店履歴から，どの店舗の何曜日に何回利用があったか集計した結果により来店傾向の類型を作成して顧客を分別するといった方法で，分別のあるデータの分別を可能にしていきたいと考えています」

同様の方法を用いて，介護保険制度の過去3年間の要介護認定結果から高齢者の心身状態を類型化し，このデータと要介護認定調査結果から，サービス給付の標準を与えるための高齢者分類も提案している（図1）。

関氏は，これら多量のデータを扱う作業にS-PLUSを利用している。

「使い慣れたS言語はなくてはならないもの。オブジェクト指向に基づいて関数を定義でき，数十万件規模のデータも処理できる信頼性があります。何を解析したいのかという目的に合わせて的確に処理されるため，新たな解析の手法など本来の作業に集中することもできます」

と語る関氏。また，分析結果を可視化するために必要な図を自由に描けるS-PLUSのグラフィック機能を活用し，分析結果を解釈するのに役立てている。