統計解析環境 R への機能拡張について

データ科学研究系　計算機統計グループ
教授　　中野　純司

1　はじめに

統計解析環境 R (`http://www.r-project.org/`) は、統計計算とグラフィックスの機能が豊富なフリーのプログラミング環境である。統計学研究者のコミュニティでは標準的なソフトウェアのひとつになっており、統計学を応用する他分野の研究者の間でも利用者が急速に増加している。

R は 1970 年代から開発されてきた統計解析システム S を模したものである。S の初期設計が優れていたため、当時とは激変している現在の計算機環境や技術を R はうまく吸収し続けている。そして世界中に分散している R 開発者の多大な努力により、現在の最先端の統計解析環境となっている。さらに R には、新しい機能を非常に簡単にインストールするための “パッケージ” の機能があり、これを利用して、非常に多くの機能が付加され続けている。

統計数理研究所においても、統計科学技術センターおよび情報・システム研究機構の新領域融合研究センターの活動のひとつとして、主としてスーパーコンピュータ上の R をサポートしている。本発表では、それらのいくつかの成果について述べる。

2　遺伝研 DDBJ の利用

国立遺伝学研究所では、DDBJ (DNA Data Bank of Japan) を、欧州の EBI/EMBL-Bank および 米国の NCBI/GenBank と共に、『国際塩基配列データベース (INSD)』として構築・維持・配布している。さらに Web API で提供しているいくつかの相同性検索サービスも提供しており、それを R から SOAP (Simple Object Access Protocol) を用いて直接利用できるようにした。

3　Huge TLB 機能の利用

現代の CPU は TLB (Translation Lookaside Buffer) を持つが、これは仮想記憶領域から実記憶領域へのハードウェアキャッシュ(4-8KB 程度) である。最近の OS では、メモリを利用して Huge TLB 機能 (2-256MB 程度) を実現しているものが多い。この機能は、Linux と Windows では Huge TLB、Sun Solaris では Intimate Shared Memory (ISM)、AIX では Large Page などと呼ばれている。この機能は大量のデータを扱うときの “TLB miss” を防ぐのに有効であるので、R からこれが利用できるようにした。

4　Grid 用並列計算インターフェース

国立情報学研究所が中心となった NAREGI（National Research Grid Initiative）プロジェクトは最先端の研究や教育活動を支援するための学術グリッドを構築することを目的とするものであり、2008 年に NAREGI ミドルウェア Ver.1 をリリースした。これは離れた多数のスーパーコン

ピュータを並列利用する技術であるが、かなり複雑な構成を取るものである。われわれはその中の機能である Grid RPC を利用して R の グリッド化を行った。

5　パーソナルクラスター用並列計算サポート機能

現在では、規模は劣るがスーパーコンピュータと類似の並列計算機能をパソコンのクラスターでも実現できる。われわれが主としてスーパーコンピュータのために開発してきた R の並列計算のための機能もこのようなパーソナルクラスターでも利用できる。ただし、そのためには多くのフリーソフトウェアの準備が必要である。これらのフリーソフトウェアのダウンロード、コンパイルなどは面倒な上、しばしば R 用に変更を加える必要がある。そこでわれわれは Debian/GNU Linux 上でそれらを簡単に実現するためのパッケージおよび “ヘルパー” を作成した。

以上は、中間栄治氏との共同研究である。

6　CATDAP のパッケージ化

CATDAP は坂元慶行氏等により開発されたカテゴリカルデータのモデル分析のためのソフトウェアである。今回、これを R のパッケージとして、より利用しやすくした。これは嵯峨優美氏との共同研究である（図 1 ）。

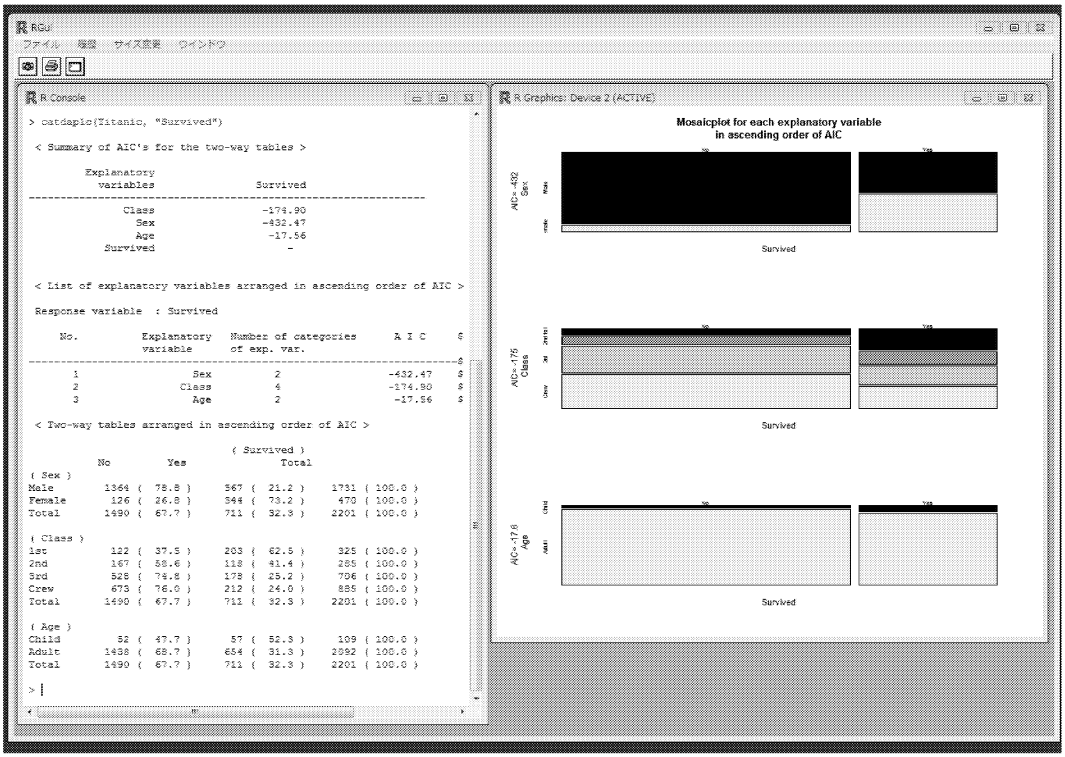

図 1: CATDAP on R