

第 6 章

結論

本研究では、並列計算機の計算パワーにより、高速で精密な統計データ解析を行い、データの持つ特徴を自動的に抽出することを目的とした。また、様々な環境で一般ユーザが使用出来ることを目的とした。

第 2 章では、開発を行った統計データ解析手法の内容と、その並列プログラムについて述べた。統計データ解析手法には、大きく分けて目的変量を持つものと持たないものがある。また、その目的変量が数値データか質的データか、目的変量を導き出すための説明変量が数値データか質的データかによっても用いる手法が違う。本研究では、基本統計量、重回帰分析、主成分分析、判別分析、数量化理論 I 類、数量化理論 II 類、クラスタ分析のプログラムを開発したので、これらによりデータ形式により使い分けるべき手法が一通り網羅出来たと言える。

また、並列計算機の高速性を利用して、相関係数をもとめる処理を拡張し、データの区間を区切った細かい解析を行うことを可能にした。更に他の手法では、全変量の組合せにおける解析なども行えるようにした。

開発方法では、汎用的に利用出来るシステムを目指し、通信ライブラリに MPI を用いた。MPI を用いて開発したことにより、様々な並列計算機システム上での実行が可能となった。第 3 章で示したように、実際に CP-PACS, RS/6000 SP, パソコン・クラスタという異なる環境上で、プログラムを変更すること無く実行出来ることを確認した。

第 3 章では、開発した統計データ解析のプログラムを用いて、CP-PACS, RS/6000 SP, パソコン・クラスタで行った性能評価について述べた。

オーバヘッドの面から性質の違う基本統計量、重回帰分析、主成分分析を用いてデータの変量数、サンプル数を変えながら測定を行った。

その結果、基本統計量では PVP-SW 機構により高速なベクトル演算が出来、ハイパクロスバ構造でデータ転送の高速な CP-PACS が一番有利であるという結果が得られた。

実行速度の比は、 $P=8, M=100, N=50000$ の時、パソコン・クラスタ:CP-PACS:RS/6000 SP = 1 : 14.041 : 1.154 となった。

重回帰分析では、ブロードキャスト処理により頻繁にデータのやりとりをするため、通信速度の遅いパソコン・クラスタでは非常に時間がかかり、実用的な時間では処理出来なかった。また、本来一番高速な通信能力を持っているはずの CP-PACS が、MPI を使い機種固有のチューニングを行わなかったためかスペック通りの性能を発揮出来なかった。よって、重回帰分析では RS/6000 SP が一番適しているという結果が得られた。

実行速度の比は、 $P=8, M=100, N=50000$ の時、パソコン・クラスタ:CP-PACS:RS/6000 SP = 1 : 9.180 : 10.330 となった。

主成分分析では、処理の途中に通信処理を行っていない。よって、計算の速度のみが全体の処理速度に影響すると考えられる。しかし、本研究では、コンパイラの関係で CP-PACS の PVP-SW 機構が使い切れていない部分があり、RS/6000 SP の方が高速に処理出来るという結果が得られた。また、通信処理が無いおかげで PU 数の増加による速度低下が小さく、パソコン・クラスタでも実用的な時間で処理出来ることが分かった。

実行速度の比は、 $P=8$, $M=100$, $N=50000$ の時、パソコン・クラスタ:CP-PACS:RS/6000 SP = 1 : 3.204 : 3.984 となった。

第 4 章では、システムを利用するために行った環境構築について述べた。

遠隔地から並列計算機システムを操作するインタフェースを、汎用的な利用を考慮し、Java 及び HORB を用いて構築した。また、解析結果の理解を助けるための 3 次元グラフ表示を、汎用的な利用を考慮し、VRML を用いて行った。

以上により、遠隔地にある様々な環境のシステムからの統計データ解析システムの利用が可能となった。また、コンピュータや統計解析の専門家ではないユーザの利用や理解を助けられるようになった。

第 5 章では、実際のデータ (Human Development Report) を用いて行った解析の結果について述べた。

解析の結果、これまでの手法でも求められたような一般的な結果とともに、以下に示すようなものを一例として、特徴的な結果もいくつか得られた。

相関係数を求める処理では、GDP と国家の全支出に占める教育費の割合との間に、区間によって正相関と負相関が現れる部分があるということを、自動的に発見することが出来た。

重回帰分析を行った結果からは、HDI という様々な指標から導き出す指標が、余命、教育、GDP という指標によりほぼ説明出来るということを、自動的に発見することが出来た。

その他にも様々な解析結果が得られ、このシステムの有用性が示された。

以上により、並列計算機の計算パワーにより高速な統計データ解析を行い、データの持つ特徴を自動的に抽出することが可能になった。

また、MPI, Java, HORB, VRML などを用いて開発することにより、汎用性を確保することが出来た。

更に、操作インタフェースの開発及び 3 次元グラフによる結果表示を行うことにより、一般ユーザの使用を容易にした。