第 1 章

序論

1.1 近年のデータ解析

近年, 情報処理機器の発展に伴い社会のあらゆる場所で大量のデータが蓄積されるようになってきており, その解析が重要になってきている.
データの集まりを捉え, 一つ一つのバラバラなデータからは得られない, データの集団が持つ様々な特徴 (法則) を眼に見えるような形で捉えようとするのが最近のデータ解析の特徴である. [1], [2]
また, このような考え方に基づいて, データベースの分野では, データマイニングという, データ集団の特徴を抽出する研究も盛んに行われている. [3], [4], [5]
しかし, 大量のデータに対して, データのあらゆる組合せについての解析のような精密な解析を行おうとすると, 膨大な時間がかかってしまう. そこで, 近年, より一般化してきた並列計算機のような高度な計算機システムを使用することが期待されている.

1.2 これまでの統計データ解析の問題点

これまでの統計解析処理における問題点として, 統計解析に不慣れなユーザは不十分な処理しか出来ない可能性があった. つまり, より正確な解析を行うためには, 統計解析や解析しようとする現象 (データ) そのものに関する知識が必要であったということである.
一方, 統計解析や解析したい現象に関しての豊富な知識を擁していた場合には, ある程度結果を予測しながら処理を行っていた. これは, 処理時間の関係などからやむを得ない面もあったが, 思いがけないような解析結果を見落としている可能性があった.

これまでに, 統計家がプログラミングに関する知識や技術を持たずに統計解析プログラムを作ることを支援するシステムの研究がなされている. [6] また, 統計学やデータ解析に不慣れな初心者のユーザが, 手法の誤用などをおこさないように, 知識ベースや推論機構を利用して解析手法の選択を支援するシステムの研究がなされている. [7], [8], [9]
しかし, このような場合でも, 有用な結果の見落としを起こす可能性は残っている.

1.3 本研究の目的

本研究では, 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う事を目的とした. これまでの統計データ解析では, 一回当たりの計算量はそれほど多くは無いと思われていた. しかし, データの変量などの全ての組合せについての解析を行う事を考慮すると計算量は膨大なものになる.
そこで, このようなこれまでは行えなかった解析を, 並列計算機のパワーによって行うことによ

り. これまでは見逃していたようなデータが持つ特徴を, 人間が手間をかけずに (自動的に) 発見出来るようになる事を目的とした.
このような処理が可能になる事により, ユーザ側の知識不足や予見によらない解析が行えるようになると考えた.

また, このシステムは, 色々な場で広く使われることを目的とした. 様々な環境での使用を可能にするために, 出来るだけ汎用的なシステムの開発を目指した.

更に, 統計解析に不慣れなユーザが使用する場合を考慮して, 扱いやすいインタフェースの開発や 3 次元グラフ表示による結果の提示などを行うこととした.

1.4 論文の構成

以下, 第 2 章では, 汎用的なシステムを目指し通信ライブラリに MPI(Message Passing Interface) を用いて開発を行った統計データ解析プログラムについて, 第 3 章では, 統計データ解析における様々な並列計算機システム環境の性能評価について, 第 4 章では並列計算機の遠隔地からの使用やユーザの使い勝手の向上, 結果の理解を目的として作成したインタフェースについて, 第 5 章では, 実際のデータの解析例について, 第 6 章では結論について, それぞれ述べる.