数理解析研究所講究録
986巻 1997年 174-180

数式処理システムとデータベースの結合とデータ解析への応用

愛媛大学 工学部　白石 啓一 (Kei-ichi Shiraishi)
愛媛大学 工学部　甲斐 博 (Hiroshi Kai)
上智大学 理工学部　齋藤 友克 (Tomokatsu Saito)
愛媛大学 工学部　野田 松太郎 (Matu-Tarow Noda)

1. はじめに

現在の情報社会では、計算結果や観測結果などから得られるデータが大量に蓄積されつつある。例えば、遺伝子情報、経済データといったものがある。このようなデータの解析にはパターンマッチングによるものや、ある種の解析技法を用いて処理する必要のあるもの等様々である。これらのうち、理工学で対象とする分析のように各時点での確定的な数値結果を得ることを重視するものの他に、経済分析のように、確定的な数値結果を重視しないものもある。特に経済の分野では、多種多様な大量データを処理して意志決定に役立つ有用な情報を抽出することの重要性が高まっている。そのために、データの統計処理のための多くのプログラムやパッケージが作成されている。しかし、これらのすべては数値計算のみを用いたものであり、結果としての数式モデルをパラメータ付きで表現する等の機能は加えられていない。さらに、必要とされるモデルの最適化の技法も単に数値的に行っているのみである。

一方、近年では数式処理の利用が広まり、数値数式ハイブリッドアルゴリズムの開発等によって、上で述べた統計データの整理、モデル作成およびその最適化を数式処理の助けを借りて行う可能性が現れてきた。そこで、本研究では数式処理を用いてこれら経済分析を行う可能性について検討することを試みる。そのためには、以下の点が必要になる。

- 計算システムとして、大量データを格納したり自由に検索するためのデータベースシステムと数式処理システムの結合
- このような計算システム上で実現可能な各種の数値数式ハイブリッドアルゴリズムの構築

しかし、現状ではこれらはほとんど実現されておらず、わずかに数式処理と数値計算の結合が実現されていること、いくつかの最適化技法に関する数式処理的アルゴリズムの開発が検討され始めていること、等があるのみである。

そこで、以下ではまず計算システムとして数式処理システムとデータベースの結合を考

え、その上での簡単なアルゴリズムの実現の可能性を考える。数式処理システムとしては分散環境での使用に適する Risa/Asir を用い、データベースシステムとしてマルチメディア対応が可能なオブジェクト-リレーショナル型の Illustra を用いる。

2. データベースシステム Illustra

Illustra とは、次のような特徴を持つデータベースシステムである。

- オブジェクト-リレーショナルデータベースシステム
  画像データなどの複雑なデータに対する問い合わせ
  例) slides テーブルに保存されている画像データ picture から夕日が写っている写真の id を検索

```
select id
from slides P
where sunset(P.picture);
```

- サポートされる機能
  - base type extension
    画像データなどのユーザ定義基本型の作成
  - complex object
    複合型、集合型、参照型などのオブジェクトの作成
  - inheritance
    他のデータ型を基にした新しいデータ型の作成と、基のデータ型の演算子、関数などの継承
  - production rule system
    リレーショナルデータベースのトリガ、アラータなどを集約、拡張したルールを扱うシステム
- データブレードモジュールによる拡張性
  - ユーザ定義のデータ型、演算子、サーバ関数、アクセスメソッドをパッケージ化したクラスライブラリ
  - 容易かつ効率的な Illustra の拡張

Illustra のようなオブジェクト-リレーショナルデータベースシステムでは、画像データの拡大や重ね合わせ等の処理も容易であり、数式等の蓄積への対応も簡単である。さらにデータブレードの活用によって、既存のリレーショナルデータベースでの SQL 以上の効果を発揮することが可能になる。

3. データベースとの結合

一般に Risa/Asir のような数式処理システムと Illustra のようなデータベースシステムを結合することによって実現すべき機能は、

- データベースに蓄積されたデータの自由な取り出し

- 処理結果のデータベースへの蓄積

等がある。これらは結局は

数式処理システムでのデータベースシステムへの問い合わせ文 (SQL) の自動発行

にあるといえる。今の場合は select 文を Risa/Asir で自動的に発行し、Illustra へ伝えることによって必要な機能を実現すればよいわけである。このための方策としては、一般に、Risa/Asir で作成した select 文を通信回線を経由して Illustra へ送信し、Illustra から対応データを受け取り、数式処理結果を再びデータベースに戻すという手法が考えられる。しかし、この場合には、通信回線を経由して、膨大な量のデータをデータベースから数式処理システムへ転送する必要がある。一般には、この点での非効率性が存在する。

一方、Illustra には、既に述べたようにデータブレードによる拡張があり、データブレードとして Risa/Asir を登録することが出来れば、膨大なデータが通信回線を流れるという点を解消することが可能になると考えられる。このような方向で数式処理システムとデータベースシステムを結合し、それらの性能評価を行い、最適な両者の結合形態を探ることが本研究の目的の一つである。しかし、残念ながら、現時点では我々の開発はここまで至っていない。そこで、ユーザが作成した問い合わせ文をデータベースへ送り、データをファイルへ書き出し、それを Risa/Asir が読み出すというシステムを作成した。以下ではファイルからデータを読み出し、関数近似を行う。

4. データの関数近似

経済分析を行う上で、統計データの整理、モデル作成およびその最適化が必要となる。ここでは今後の実際的な応用の準備的考察として、与えられたデータの関数近似を行い、今後、データ解析を行うためのアルゴリズムをインプリメントする。これによって、数式処理システムがデータ解析のツールとして有用になることを示す。データの関数近似には、様々な方法があるが、ここでは、今後より大量のデータを扱う場合の出発点として、非常に簡単なハイブリッド有理関数近似と直線回帰分析を取り上げ、読み出したデータの関数近似を行い、数式モデルを作成することを試みる。まず、ここで用いた直線回帰分析の手法を述べておく。

直線回帰分析

ここでの直線回帰分析は、モデルとして直線

$$y_i = a_0 + a_1 x_i \quad (i = 1, 2, \ldots, n)$$

を仮定し、データ系列

$$\{x_1, x_2, \ldots, x_n\}$$

$$\{y_1, y_2, \ldots, y_n\}$$

を用いて係数を推定する。その方法は、観測値を推定値との残差の 2 乗和についての最小 2 乗法による。即ち、時点 i における推定値を y_i^* とすると、

$$S = \sum_{i=1}^{n} (y_i^* - y_i)^2$$

月	売上 (十万円)	人口 (千人)	月	売上 (十万円)	人口 (千人)
1	8.2	6.0	13	14.6	7.7
2	8.4	6.2	14	14.8	7.8
3	9.0	6.4	15	16.0	7.8
4	9.4	6.6	16	16.0	7.8
5	9.6	6.8	17	18.0	7.8
6	10.4	7.0	18	20.0	7.9
7	10.8	7.1	19	21.2	7.9
8	11.2	7.2	20	22.0	8.0
9	12.0	7.3	21	24.0	8.2
10	12.8	7.4	22	26.8	8.4
11	13.4	7.5	23	28.8	8.6
12	14.4	7.6	24	30.0	8.8

表 1. コンビニエンスストアの月次パン売上金額と団地の人口

を最小にするように係数 (a_0, a_1) を決定する。S を変数 (a_0, a_1) について偏微分し、各々の結果を 0 とおいた連立方程式を解くことで決定できる。

$$\frac{\partial}{\partial a_0}\sum_{i=1}^{n}\{y_i-(a_0+a_1x_i)\}^2=-2\sum_{i=1}^{n}(y_i-a_1x_i-a_0)=0$$

$$\frac{\partial}{\partial a_1}\sum_{i=1}^{n}\{y_i-(a_0+a_1x_i)\}^2=-2\sum_{i-1}^{n}(x_iy_i-a_1x_i^2-a_0x_i)=0$$

この例は非常に簡単であるが、高次の関数近似への適用も考え、ガウス消去法を Asir のユーザ言語で作成し、連立方程式を解いた。

例 1

表 1. はあるコンビニエンスストアの月次のパンの売上金額 (十万円単位) と団地の人口 (千人単位) の 2 年間のデータである。このデータをもとに、続く半年の月次のパンの売上予測を行う。

月を x、パンの売上を y とする。

- ハイブリッド有理関数近似

  このデータをもとに、ハイブリッド有理関数近似を行うと、次の式が得られる。

$$\frac{-534.375x+33416.6}{4.65534x^2-256.696x+4196.29}$$

  ここで、データの x_i に関しては、$[0,1]$ に収まるように 1/24 倍した後、ハイブリッド有理関数近似手続きを適用し、得られた式の x を $x/24$ で数式処理的に置き換えることにより、上式を得た。

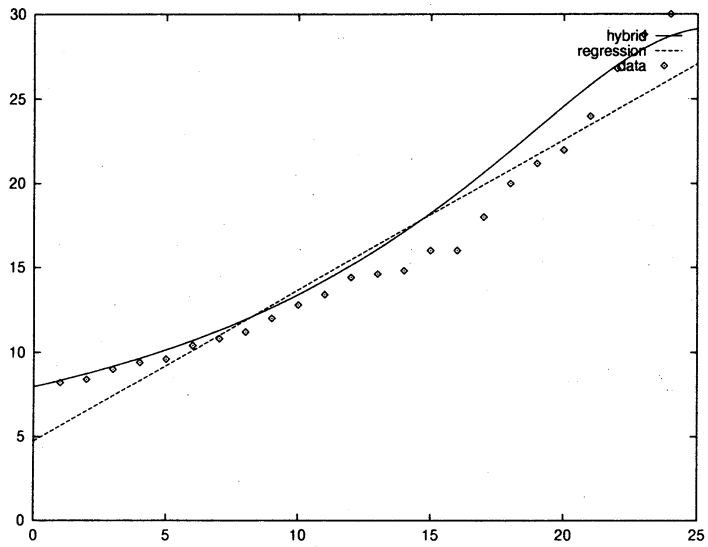

図 1. 月次パン売上金額とその関数近似

	残差の 2 乗和
ハイブリッド有理関数近似	55.1949
線形回帰分析	79.2184

表 2. 例 1 における残差の 2 乗和

- 直線回帰分析
  直線回帰分析では、連立方程式

$$24a_0 + 300a_1 - 381.8 = 0$$

$$300a_0 + 4900a_1 - 5799.2 = 0$$

を解く。結果として、次の式が得られる。

$$0.892783x + 4.74855$$

これらの式のグラフを図 1. に、この時の残差の 2 乗和を表 2. に示す。

ハイブリッド有理関数近似は、このようにばらつきの多いデータに対しても残差の 2 乗和という視点からは、比較的良好な結果を与えることが判る。

例 2

表 1. のデータをもとに、団地の人口による月次のパンの売上予測を行う。
人口を x 、売上を y とする。

- ハイブリッド有理関数近似
  このデータをもとに、ハイブリッド有理関数近似を行うと、次の式が得られる。

$$\frac{228.409x^3 - 3660.26x^2 + 15710.2x - 6184.38}{-6.86281x^3 + 275.341x^2 - 3128.1x + 11025.4}$$

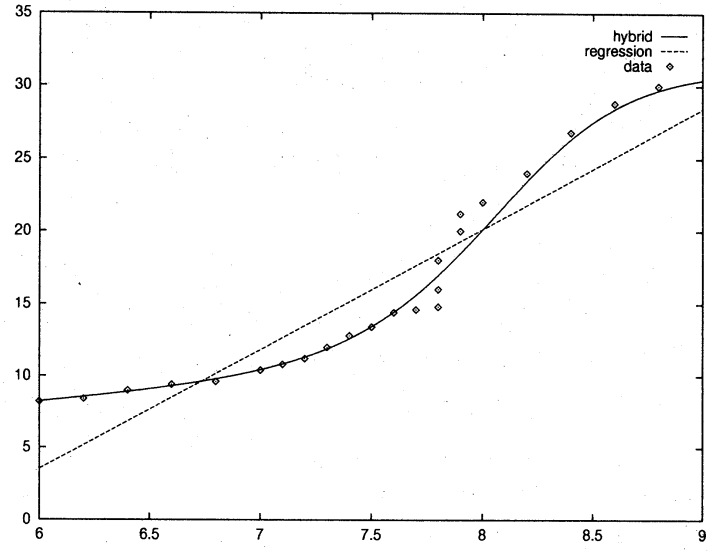

図 2. 団地の人口による月次パン売上金額と
その関数近似

ここでも、例 1 と同様、データの x_i を $(x_i-6)/3$ と変換し、$[0,1]$ の量にした後、ハイブリッド有理関数近似手続きを適用し、得られた式の x を $(x-6)/3$ で数式処理により置き換え、上式を得た。

なお、このデータはそのままでは、14 月から 17 月に $x_i = 7.8$ という同じ値を持つ。そのため、ハイブリッド有理関数近似を行おうとすると、補間手続き中のガウス消去が失敗する。ここでは x_i に同じ値を持つ点がないように、$x_i = 7.8$ に対し、y_i の値の平均を新しく y_i とする操作を行い、ハイブリッド有理関数近似を行った。この種の問題はデータ量の増大と共に頻繁に発生するものと思われる。より組織的な解決法を確立する必要があると思われる。

- 直線回帰分析

直線回帰分析では、連立方程式

$$24a_0 + 179.8a_1 - 381.8 = 0$$

$$179.8a_0 + 1359.18a_1 - 2961.52 = 0$$

を解くことになる。これを解くと、次の式が得られる。

$$8.30998x - 46.3472$$

これらの式のグラフを図 2. に、この時の残差の 2 乗和を表 3. に示す。

例 1 と同様、ハイブリッド関数近似は良好な結果を与える。

5. むすび

数式処理システムとデータベースによる大量データの解析システムの可能性を示し、そのための予備的なアルゴリズム実行例として、ハイブリッド有理関数近似と直線回帰分析

	残差の 2 乗和
ハイブリッド有理関数近似	22.8027
線形回帰分析	154.855

表 3. 例 2 における残差の 2 乗和

によるデータの関数近似を行った。この過程で、ハイブリッド有理関数近似は、このようなデータの近似にも有効であることが判った。しかし、ここで扱ったような小規模な問題においてさえ、そのままのデータをハイブリッド有理関数近似に与えるとデータが重複するなどうまく関数を求められない場合があることも判った。本稿ではそれらのデータの平均によって重複データを置き換えるという単純な方法を用いたが、このような重複を防ぐため組織的な手法を確立する必要がある。

今後の課題として、まず

- 数式処理システムとデータベースシステムの最適な結合方法の確立
- ユーザインタフェースの強化

がある。通信による結合を行い、問い合わせ文、データの自由な受渡しを行えるシステムにすることは当然として、大量データの受渡しを行うと通信速度の問題が出て来るので、効率化を行う必要がある。このとき、 Risa/Asir のデータブレード化も検討すべきことの一つだろう。ユーザインタフェースは、 Risa/Asir からの問い合わせ文発行を実現することで、良くなると思われる。また、

- その他のデータ解析法のインプリメント

があげられる。以上でも挙げた通り、統計データの整理、モデル作成、およびその最適化などを実現し、例えば、経済分析を行えるようにすることが重要である。

参 考 文 献

[1] Michael Stonbraker, OBJECT-RELATIONAL DBMSs, Morgan Kaufmann Publishers, Inc., San Francisco, California (1996)

[2] 甲斐博 · 野田松太郎, “ハイブリッド有理関数近似とデータの平滑化”, 日本応用数理学会論文誌 Vol.3, No.4, pp. 323–336 (1993)

[3] E. クライツィグ, 数値解析, 培風館, 東京 (1988)

[4] 杉原敏夫, 経営 · 経済のための時系列分析と予測, 税務経理協会, 東京 (1994)