ハイパフォーマンス
コンピューティング 77－27
（1999．8． 4）

表データからの傾向抽出方式と並列化手法

仲瀬 明彦、久保田 和人、酒井 浩、小柳 滋
新情報 並列応用東芝研究室

概要

データベースに蓄積されたデータから、有用な知見や法則を発見するデータマイニングの一分析手法として、関係データベースの表形式のデータからレコードの分布の傾向を発見する方式の開発を行なった。分析対象データは、関係データベース等の表形式のデータとし、与えられた条件で選択されたレコード群に共通する特徴をデータベースの属性の属性値を用いて説明する。結果として、「属性AがX以上Y以下かつ属性BがZの時、レコードが与えられた条件を満たす確率はX%である。」等の規則が生成される。本分析手法は、分析対象属性の全ての組合せについて分析を行なうので、一つの組合せを一つのタスクとすることにより、並列化が容易である。超並列マシン上で2種類の負荷分散手法で本データマイニングアルゴリズムを並列化し、分析対象データの種類によって並列効果の違いが生じることを確認した。

Association Analysis from Table Data and its Parallel Execution

Akihiko Nakase, Kazuto Kubota, Hiroshi Sakai, Shigeru Oyanagi
RWCP Parallel Application Toshiba Laboratory

We developed a data mining system which automatically find association rules from table data. In our system, a interested attribute can be explained using target attributes. We also made parallel execution version of this mining system and proposed two types of load balancing strategies. We compared scalability of these strategies with two types of data on massively parallel machine, and analyzed bottle-neck of each method.

1 はじめに

近年、情報処理技術やネットワークインフラの普及により、多種類の情報を電子的に採取・記録する技術が急速に発達し、また、外部記憶装置の大容量化・低価格により、大規模なデータベースを構築することが可能になってきた。

データウェアハウスは、このような大規模データベースを有効に利用することを目指したもので、長年蓄積したデータベース内のデータや、企業内の部所間での横断的なデータベースを統合的に利用することを可能にする。データウェアハウスの利用技術としては、データ内から隠された傾向や知識を発見するデータマイニングが注目されている。

我々は、関係データベースにおいて、一つのリレーションのN個の属性からN次元の空間を作り、N次元の空間内での各レコードの散布状況や注目する属性の属性値の分布状況から、N次元の空間内での特徴的な部分空間を発見するスライスアンドダイス分析手法を開発した。

表形式のデータのN個の属性からN次元空間でのレコードの分布の傾向を発見するアルゴリズムは、Subspace Clustering [1], 特徴的ルール生成 [2], Region Splitting [3] 等として報告されているが、一般的に分析対象データの属性数が大きくなると膨大な計算時間を必要とするので、並列処理により高速化することが望ましい。

本稿では、スライスアンドダイス分析アルゴリズムの概要と、本アルゴリズムを超並列機上で並列実行した時の評価について述べる。

2 スライスアンドダイス分析

2.1 概要

関係データベースなどの表形式のデータにおいて、一つのリレーションのN個の属性からN次元の空間を作り、N次元の空間内での各レコードの散布状況や注目する属性の属性値の分布状況から、N次元の空間内での特徴的な部分空間を発見することは、有用なデータ分析手法である。

我々は部分空間内の2つの特徴に注目する。

(1) 正例負例の判別:
分析対象のレコードを与えられた条件で正例と負例に区別した時、正例の密度が高い部分空間を発見する。

(2) 注目属性の値の傾向:
ある属性に着目して、その属性値が全体の平均に比べて大きい値を多く含む部分空間を発見する。

(1) の例を示す。

表1に示す小売業の売上に関するデータベースで、ある廃棄ロスのある商品は商品分類と売上の季節から説明付けられるか、単価と売上数がどのような値であれば廃棄ロス数が多い例が多いか等を発見すると、それらの傾向や知見を仕入れ数量の決定や品揃えなどのビジネス戦略に利用することが可能となる。

表 1: 小売業の売上データベースの例

商品名	単価	売上数	廃棄数	利益率
梅にぎり	105円	153個	6個	4%
鮭弁当	450円	35個	24個	10%
新鮮牛乳	155円	241個	0個	9%
プリン	150円	25個	0個	2%
.....	.....	.....		

以下に廃棄ロス数が0ではない商品を特徴付ける規則を発見する例を示す。

表1のデータにおいて、廃棄ロス数が0より大きいレコードを正例、廃棄ロスが0以下のレコードを負例とする。

次に、表1の例で単価と売上数の属性を分析対象属性とし、2次元の空間を生成する。

この空間内で、各レコードを対応する座標に記録する。その際に、レコードが正例であったか負例であったかも記録し、図1のような空間を生成する。(図1は、表1のようなデータベースでレコード数が多くなったものを仮定した例である。)

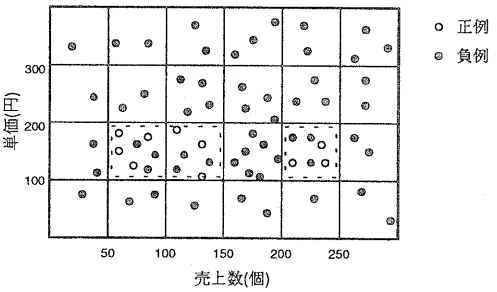

図 1: 正例と負例の分布の図

例えば、図1の例では、正例レコードが10個、負例レコードが60個あるので、正例レコードの全体に対する割合は約14パーセントである。

ここで、売上数が50個～150個かつ価格が100円～200円の区画では全13レコード中正例レコードが7個あり (約54パーセント)、売上数が200個～250個かつ価格が100円～200円の区画には全6

レコード中正例レコードが3個ある(約50パーセント)ので、正例レコードが密集している副空間と認識する。

結果として、「売上数が50個～150個かつ価格が100円～200円の商品は、廃棄ロスの出る可能性が多い。」、「売上数が200個～250個かつ価格が100円～200円の商品は廃棄ロスの出る可能性が多い。」等の規則が生成される。

次に(2)の例を示す。

表1の例で利益率の属性に注目し、売上数と単価を軸とする2次元空間内に、各レコードの利益率の値を記録する。(図2)

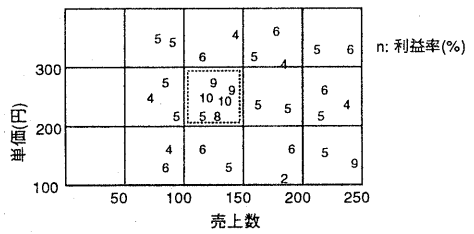

図2: 注目属性の値に注目する場合

図2からは、利益率の全体の平均が6%であるが、売上数100個以上150個以下、単価200円以上300円以下の商品では、利益率の平均が9%と他の部分に比べて利益率が大幅に高いことが分かる。

このように、関係データベースのn個の分析対象属性でn次元の空間を作り、各レコードが与えられた条件を満たすか否かの記述や、注目する属性の値を記録した多次元の集計表で、特徴的な副空間を発見する分析方法をスライスアンドダイス分析と名付けた。

2.2 スライスアンドダイス分析の逐次アルゴリズム

スライスアンドダイス分析は、以下のステップで実行される。

1. 分析対象属性、正例条件、注目属性の指定
2. 多次元集計表の生成
3. 特徴的な副空間の発見
4. 特徴的な副空間の拡張
5. 結果の出力

以下に各ステップの詳細を述べる。

2.2.1 分析対象属性、正例条件、注目属性の指定

対象データから、分析対象としたい属性群(各属性を軸とした空間を作りたい属性群)を分析対象属性として指定する。

対象データのレコードを正例と負例に分け、正例と負例の分布の傾向を分析する場合は、正例条件となる条件節[1]を指定する。

ある属性に注目して値の傾向を分析したい場合は、注目属性を1つ指定する。

2.2.2 多次元集計表の生成

分析対象属性のN個の属性からN次元の空間を作成する。以下、このN次元空間を多次元集計表又はデータキューブと呼ぶ。

図1のような区画を形成するには、データキューブの軸となる分析対象属性の各属性を区画に区切らなければならない。この区画をメッシュと呼ぶ。

データキューブの各軸のメッシュの区切り方は、分析対象属性の型によって、以下の2通りの方法を用いる。(図3)

- 数値型の属性

  ある数値属性の上端と下端の2.5%程度の値をはずれ値として除外した後、最大値と最小値の間をk等分したメッシュ区切りとする。kはシステム固定値[2]、又はユーザが与える。

- 文字列型の属性

  1種類の文字列を1つのメッシュに入れる。

分析対象のn属性でn次元の空間を作る時、各属性のメッシュ数を N1,N2,N3,,,Nn とすると、データキューブには $S=N1 \times N2 \times N3 \times \ldots \times Nn$ 個の区画が出来る。

ここで、分析対象テーブルの各レコードがどの座標のメッシュに入るか計算し、そこにレコード

[1] SQL の where 句で指定するのが一般的である

[2] あまり小さくても意味がなく、大き過ぎても結果が膨大となってしまう。20～50程度が妥当である。

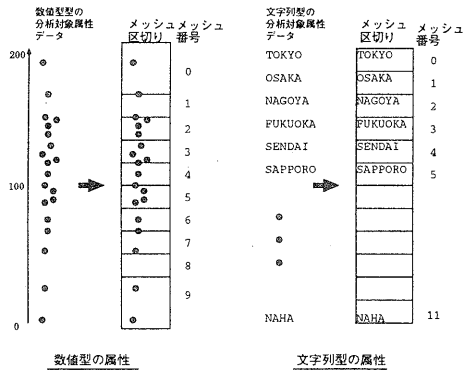

図 3: メッシュに区切る方法

数、注目属性の正例／負例の区分、注目属性の値、等を格納する。

2.2.3 特徴的な副空間の発見

データキューブの各メッシュを走査し、部分空間の判定を行なう。

判定の基準は、分析の手法によって以下のいずれかの判定方法を用いる。

- 正例負例の判別:
  - 正例レコードが、与えられた個数以上存在する副空間を判定合格とする。
  - 正例レコードと負例レコードの比率が、与えられた値以上である副空間を判定合格とする。
  - 統計的な検定手法を用いて [4]、正例レコードと負例レコードの比率を全空間内で一様と仮定した時、正例／負例レコード数の組合せが起こる確率が、与えられた値以下である副空間を判定合格とする。
- 注目属性の値の傾向:
  - 閾値 N を与え、その副空間内の注目属性の値の平均が、全体の注目属性の値の平均より N 倍以上大きい副空間を判定合格とする。
  - 統計的な検定手法を用いて [4]、注目属性の値の平均が全空間内で正規分布に従うと仮定した時、その副空間での注目属性の平均値の取る値の確率が、与えられた値以下である副空間を判定合格とする。

スライスアンドダイス分析では、上記の判定方法を変更するだけで、多種の分析に対応出来る特徴がある。

2.2.4 特徴的な副空間の拡張

2.2.3 で選ばれた特異な副空間を、超矩形の形状を保持したまま数値属性の軸方向に拡張する。(図 4)

拡張後の副空間が、2.2.3 で示した判定基準を満たしていればそれを新たな解とするする。

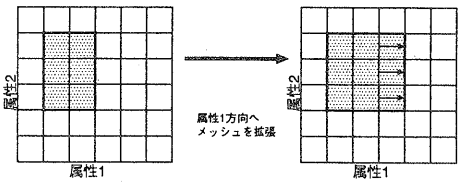

図 4: メッシュを拡張する例

2.2.5 結果の出力

以上の操作で得られた副空間の記述を前件、副空間の内容の記述を後件として、if-then 形の規則を生成する。

結果は、

「if 属性 A の値 ∈ [X,Y] かつ 属性 B の値 ∈ [U,Z]

then その区間での正例割合は P%である。」

等の形式となる。

3 並列処理による実行

3.1 並列処理方式

前節で述べたようにスライスアンドダイス分析では、分析対象属性が n 個の場合、属性の組合せが $2^n - 1$ 通りあるので、$2^n - 1$ 個の部分データキューブについて分析を行なう。

1つのデータキューブの分析を1つのサブタスクとし、これらのサブタスクを複数プロセッサに割り付ける方式とする。

分析対象のデータは各プロセッサからアクセス可能なディスク上に配置し、各プロセッサともアクセス可能なメモリ上にデータキューブを保持する。(図5)

また、k($k \leq n$)次元のデータキューブは、n次元のデータキューブを射影することにより生成する。

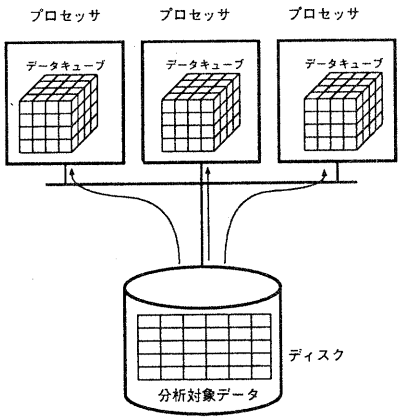

図5: 並列実行時のデータ保持方式

各サブタスクは、以下の2通りの方式で各プロセッサに割り付ける。

- 静的分割方式:

  $2^n - 1$ 個のサブタスクに0から $2^n - 2$ までの通し番号を付け、プロセッサ数をPとしてプロセッサにも0からP-1までの通し番号を付ける。

  サブタスク番号をSとすると、S番目のサブタスクはS mod P番目のプロセッサが実行する。

  データの入力と結果の出力以外はプロセッサ間の通信は無いが、サブタスクの粒度にばらつきのある場合、各プロセッサの負荷が均等にならない。

- 動的分割方式:

  サブタスクを集中管理するプロセッサを1つ設ける。その他のプロセッサは自分のタスクが無くなると管理プロセッサにタスクの要求を行ない、管理プロセッサはサブタスク番号を返す。

  サブタスクの粒度にばらつきのあっても、各プロセッサの負荷が均等になることが期待されるが、管理プロセッサが必要でありプロセッサ数が増大するとタスク要求のメッセージが管理プロセッサに集中する可能性がある。

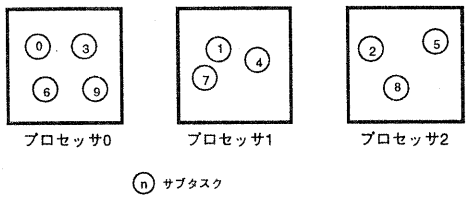

図6: 静的分割方式

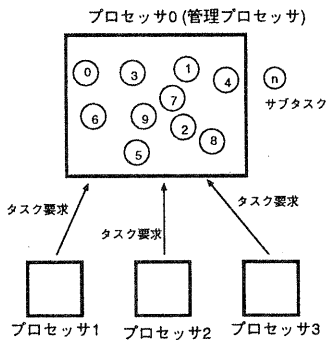

図7: 動的分割方式

3.2 超並列マシン上での並列実行

超並列マシンSR2201上でスライスアンドダイス分析の並列実行を行なった。

プログラムはSR2201上のCコンパイラでコンパイルされ、通信ライブラリはMPIを用いた。

分析データは、小売業におけるPOSデータを集計した商品の販売記録データ(約13万レコード)を用いた。

データ1では、分析対象属性を数値属性のみの14属性とし、データ2では、分析対象属性を数値属性と文字列属性の12属性とした。

上記のデータをプロセッサ数を 1 台から 250 台まで変化させ、実行時間の変化を測定した。

両データともディスク上のデータを各プロセッサのメモリ上にロードし、データキューブを作成する時間は、プロセッサ数が 1 台でも 250 台でもほぼ同じであり、全実行に占める割合は、データ 1 ではプロセッサ数 1 台の時は約 2%、プロセッサ数 250 台の時は約 75%程度である。データ 2 では、プロセッサ数 1 台の時は約 1%、プロセッサ数 250 台の時は約 65%程度である。

以下、メモリ上にデータがロードされた後の分析時間の並列効果について述べる。

データ 1 では、動的分割方式と静的分割方式で、ほぼ同等の並列効果が得られていが、プロセッサ数が多くなると、負荷分散要求の通信が集中するため、静的分割方式の方が性能が良い。(図 8)

データ 2 では、動的分割方式の方が良い台数効果が得られている。(図 9)

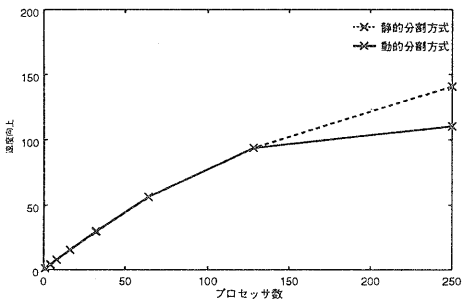

図 8: データ 1 の並列効果

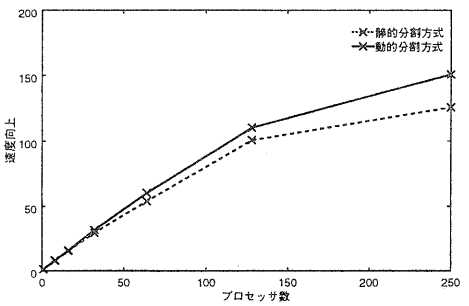

図 9: データ 2 の並列効果

データ 2 では各サブタスクの大きさにばらつきがあり、静的分割方式では負荷が均等に分割されないからである。

スライスアンドダイス分析では、分析軸に多種類の属性値を持つ文字列属性が入ると、各データキューブの大きさにばらつきが生じてしまう。

最適な負荷分散方式を事前に決定するには、各軸のメッシュ分割数を予め調べ、ばらつきがあれば動的分割方式、ばらつきがなければ静的分割方式とすることが望ましい。

4 おわりに

関係データベースの表形式のデータから、与えられた条件を満たすレコードの散布状況や注目する属性の傾向を発見するスライスアンドダイス分析とその並列実行について述べた。今後は、PC クラスタの様な分散環境で、データがディスク上に分散された状況での効率的な分析手法等を開発する予定である。

参考文献

[1] R.Agrawal, J.Gehrke, D.Gunopulos, and P.Raghaven. *"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications"*, In Proc. of 1998 ACM SIGMOD, 1998.

[2] 森田, 前田, "データマイニングサーバの開発と移動体通信業界への適用" 信学技報 *DE98-19*, 1998-10

[3] Y.Morimoto, H.Ishii, S.Morishita, *"Efficient Construction of Regression Trees with Range and Region Splitting"* , in Proc. of the 23rd VLDB Conference, 1997

[4] 岡太, 都築, 山口, "データ分析のための統計入門", 共立出版株式会社, 1995