

省メモリ版相関ルール抽出エンジンの性能評価

1 T-4

小幡 康, 三石 彰純[†][†]三菱電機（株）情報技術総合研究所

1 はじめに

現在開発を進めているデータマイニングシステムKnodiasにおいて、少ないメモリ領域内で相関ルールを抽出する“小ロット方式”のマイニングアルゴリズムによるエンジンを開発してきた[1,2]

今回、小ロット方式のマイニングエンジンの性能を最小支持度、データの規模といった様々な観点から評価するため、乱数による合成データを用いて処理速度、使用メモリ量の測定を行った。その結果、小ロット方式は従来方式に比べて優れた性能を示すことが分かり、データの規模に関しては、レコード数とレコード長が実行時間に与える影響が確かめられた。本稿では、この評価について報告する。

2. 評価の概要

本評価では、マイニング対象データを合成データ生成ツールを用いて作った。本ツールのデータ生成手順は以下の通りである。

- ①指定数の属性からなる表形式のデータベースを想定し、属性名と属性値からアイテムを作る。
- ②相関ルールの候補となるアイテムセットを指定数生成する。
- ③候補アイテムセットを繋ぎ合わせてレコードを指定数作る。

評価用のデータは実データに近い性質を持つことが望ましいので、健康診断の実データを基にして、抽出されるルール数が同等となる様にツールのパラメータを決定した。

この評価はWindowsNT4.0が稼動するApricot FT8000(プロセッサ:Pentium Pro 200MHz、メモリ:128MB)上で実施し、小ロット方式と従来方式による2つのエンジンについて、相関係数による正の

相関ルール抽出の処理時間とメモリ量を測定した。

3 評価結果

3.1 最小支持度による従来方式との性能比較

レコード数100k、平均レコード長10のデータの、最小支持度による処理時間、メモリ量の測定結果を図1、図2に示す。

最小支持度0.05%以下の領域では、全アイテムセットが仮想メモリ空間を超えるため、従来方式では相関ルール抽出が不可能である。

最小支持度0.05%~0.5%の領域では全アイテムセットが実メモリ量を超えるが、スワップを起こさない小ロット方式の方が実行速度は速い。

最小支持度が0.5%~0.2%の領域では、アイテムセットを格納するハッシュ木のメモリ管理の最適化が可能であることから小ロット方式が速い。

最小支持度が0.2%以上になると全アイテムセットが実メモリ内に収まるため、小ロット方式では一時ファイルの入出力処理の負荷がある分、従来方式の方が処理時間が短くなっている。

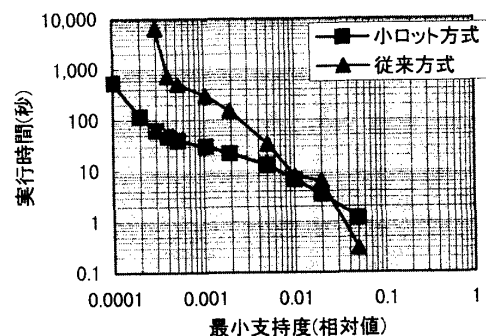


図1 最小支持度と実行時間

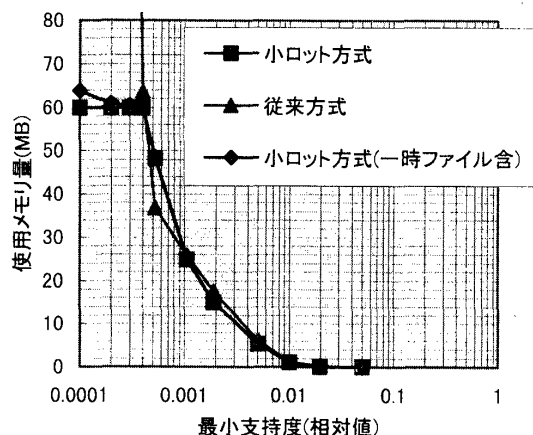


図2 最小支持度と使用メモリ量

メモリ量に関しては、従来方式では最小支持度が小さくなるに従って際限なく増加するが、小ロット方式ではアイテムセットを分割してメモリに割り当てているため、60MBに達した後は一定値を保つ。また、小ロット方式におけるラージアイテムセットを格納する一時ファイルの大きさを加えたメモリ量と従来方式のメモリ量を比較してみても、従来方式の方がはるかに多くの記憶領域を必要としていることが分かる。

3.2 データの規模と実行時間の関係

対象データのレコード長、レコード数を変えて、実行時間の変化を測定した。最小支持度は1%とした。図4、図5にその結果を示す。

レコード長と実行時間の関係では、レコード長が大きいほど、多くのアイテムセットが対象データ中に内在するため、実行時間は急激に変化する。また、一時ファイルの大きさ(アイテムセット数に比例)は実行時間と比例に近い関係であることが分かる。アイテムセット数については、レコードより生成される組み合わせの数に依存すると考えられる。レコード長 n の場合、長さ k のアイテムセットの数は最大で nC_k のオーダーとなり、実際のアイテムセット数もこの値に近い増加傾向を示すと考えられる。

また、レコード数に関する推移ではアイテムセット数がほぼ一定であり、実行時間とレコード数は線形に近い関係であることが分かる。

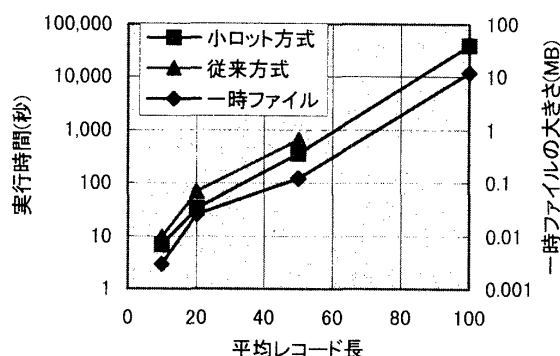


図3 レコード長と実行時間(レコード数100k)

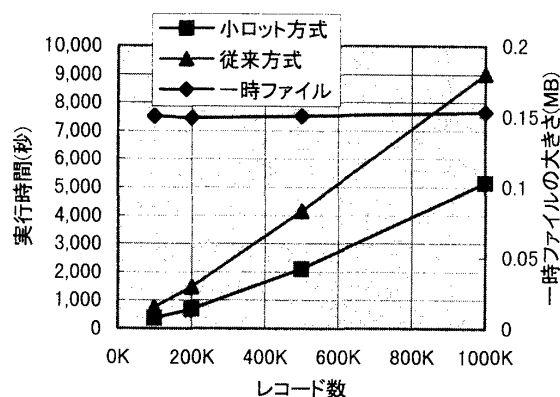


図4 レコード数と実行時間(レコード長50)

4 おわりに

小ロット方式による相関ルール抽出エンジンが、従来方式では不可能であった領域でもマイニングが可能であること、従来方式でマイニング可能な領域では従来方式と同等以上の性能が得られることを示した。また、データ規模の拡大による実行時間の変化傾向についても確かめ、相関ルール抽出の実行時間は、レコード数と線形に近い関係があるが、レコード長に対しては急激な変化を示すことが分かった。

今後はこの評価結果を基に、実行時間が大幅にかかる領域における性能を予測するための解析モデルを設定することが課題である。

参考文献

- [1] 小幡, 他: 有限メモリ空間で相関ルールを抽出するマイニングアルゴリズム, 第56回情報処全国大会1998
- [2] 三石, 他: 省メモリ版相関ルール抽出エンジンの開発, 第58回情報処全国大会1999