データマイニングにおける時系列データの処理法

5E－2

田島 玲　沼尾 雅之
日本アイ・ビー・エム株式会社 東京基礎研究所

1　はじめに

データベースからの知識獲得の研究は、様々な手法、応用分野について広く行なわれているが、その多くがアルゴリズムに主眼をおいており、例題を取り上げる際、問題を単純化するだけでなく、そのデータの規模をも縮小してしまっている。しかし、実際のデータ、例えば、企業の持つ顧客、購買データベースなどは数百万レコード、数百 MB-数 GB にも及ぶ。こうした大規模のデータを対象として、現実的な実行時間で処理を行なうには、現時点では、機能が単純なものに限定されているが大量のデータを効率良く扱えるエンジン ([1],[2]) を用いる必要がある。

こうしたエンジンは一般的に、入力データのバリエーション、フォーマットも限定されている。そこで方法論としては、マイニングエンジンにあらゆるデータを放り込む、という形は不可能であり、様々な種類のデータ間の相関関係を、試行錯誤のなかで多角的に分析する、という形となるのが現状である。

ここで、一回の試行は、(1) データベース中のデータの一部を選択してそれに対し様々な形で前処理を施し、(2) エンジンにかけて解析し、(3) 後処理をして分析する、という手順からなる。つまり、効率良く意義のある結果を得るためには、エンジンだけではなく、その前処理において、いかに有効な情報を織り込むかが重要となる。

本稿では、時系列データの前処理の方法をとりあげる。利用するエンジンの入力としては、例えば流通系企業のデータを例にとると、顧客情報、商品情報、POS 情報などが考えられるが、この中で、連続的な値をとる時間データはそのままでは単なる ID としてしか機能しない。それ以上の情報を得るには、時間軸に沿った解析によりその「時間」の持つ意味を抽出する、という処理を必要とする。そこで、時系列データに対する前処理の一方法として、時間データを量子化された属性に変換することにより、他の属性情報と同等に扱う方法を検討する。

Pre-processing time-sequence data for data mining
Akira TAJIMA, Masayuki NUMAO
IBM Research, Tokyo Research Laboratory

2　生データの解釈

流通系のデータを例にとれば、顧客住所は店舗からの距離、購買日時は季節、天気、曜日、時間帯というように、存在するデータは抽出しようとする知識の語彙に応じて様々な解釈が成り立つ。これらの解釈を、地図情報、天候などの外部データベースの利用、あるいはデータの前処理により、適切に量子化し、推論エンジンへの入力データに組み込むことにより、より質の良い情報を抽出することが可能である。

3　分析に用いる概念

前述の推論エンジンの入力は以下のように、一つの TRANSACTION に複数の ITEM、という形をとる。例えば、ITEM: 商品 ID、TRANSACTION: 顧客 ID といった応用が考えられる。

$T_1 : I_{11}, I_{12}, I_{13}, I_{14}$
$T_2 : I_{21}, I_{22}, I_{23}$
$T_n : I_{n1}, I_{n2}, I_{n3}, \ldots, I_{nk_n}$

推論エンジンは、この入力から、TRANSACTION のクラスタリング ([3])、ITEM 間の相関関係の発見等を行なう。

4　“時間” の量子化

2節に挙げたように、時間データを量子化する手段は様々考えられる。実在の数百 MB に及ぶ流通業のデータを複数解析した結果、一般に図1の例のように時期により大きな変動があるという事実が得られた。そこで、本研究では、「ある ITEM について、時間軸に沿う変動の中でどの時期に起きた事象か」という視点でタイムスタンプを解釈し、量子化する方法を検討する。

ここでは、以下の方法により、タイムスタンプを EARLY,PEAK,LATE,OFF の 4 値をとる TIMING へと量子化し、推論エンジンへの入力とする。

1. ITEM を構成する各 ID ごとに、時間軸に沿い頻度のヒストグラムを得る。
2. 5節の認識アルゴリズムにより、1から $ITEM, TIMESTAMP \rightarrow TIMING$ の参照表を得る。

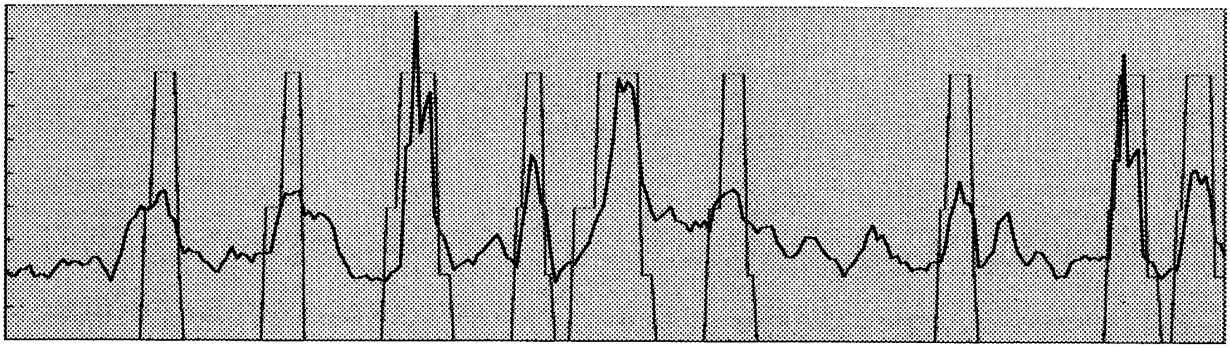
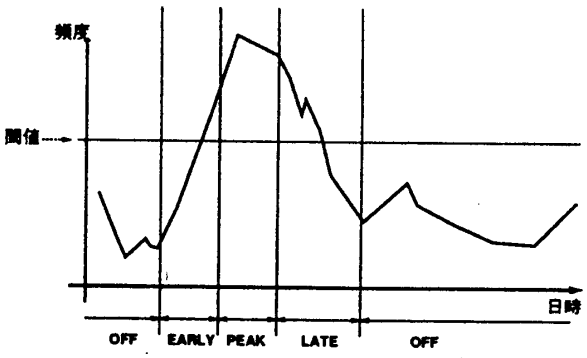


図 1: 認識結果例

3. タイムスタンプ、2の表を参照し、ITEM + TIMING を仮想 ITEM として、エンジンへの入力データを作成する。

   例えば、030001700010 という ID の ITEM は、発生時期により、EARLY-030001700010、PEAK-030001700010、LATE-030001700010、OFF-030001700010 の 4 つの仮想 ITEM に展開される。

5 認識アルゴリズム



1. 平均、標準偏差を算出し、その線形和により閾値を決定する。

2. ヒストグラム中、閾値を越えるピークを列挙する。

3. 各ピークにつき、前後の local-minimum で挟まれる区間を対象とし、最高点との間隔を一定の比率で配分して EARLY、PEAK、LATE の区間を決定する。

4. 3に該当しない区間を OFF とする。

ここで、現状ではアルゴリズム内で使用する数値パラメータは生データの特性に応じて調整している。

6 評価

ITEM+ID を仮想 ID として使用する方法では、ID の総数が 4 倍になり、それがクラスタリングやルール生成の質に影響を与えていた。時間等の連続値の量子化とともに、そうした付加的情報の入力データへの組み込み方法もさらに検討する必要がある。

7 おわりに

連続値を含み、かつ大規模なデータからの知識獲得の方法として、シンプルなマイニングエンジンと前処理を組み合わせる方法を提案した。その一例として、発生時刻の情報を持つデータに対し、時間変化を解析する前処理を施すことで時間情報の量子化を行なった。

大規模データからの知識獲得の方法論も確立していないため、本研究の結果を客観的に評価することは困難である。しかし、TIMING 情報を抜いたのみのデータとの比較から、有効性は確認された。

今後の課題としては、評価法の確立、認識アルゴリズムの強化、ドメインに適した量子化単位の発見等が挙げられる。

参考文献

[1] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules," *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1994.

[2] Chantal Bédécarrax, et al.: "A New Methodology for Systematic Exploiation of Technology Databases", *Information Processing & Management*, Vol.30, No.3, pp.407-418, 1994.

[3] 清水 周一, 木村 雅彦, 沼尾 雅之, "クラスタリング手法を用いたバスケット解析," 情報処理学会第 *51* 回全国大会予稿集,5E-03, 1995.