データマイニングにおける時系列データの処理法

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。しかし、多くの研究において、問題を単純化するだけではなく、そのデータの規模も縮小してきている。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、現実的な実行時間内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。

ここでは、一回の試行である。(1) データベース中のデータの一部を選択してそれに密度の有するデータを処理する、(2) エンジンにかけて解析し、(3) 後处理を受けて処理するという手順をとる。つまり、効率良く速度のある結果を得るためには、エンジンだけではなく、その処理において、いかに有効な情報を取り込むかが重要となる。

本稿では、時系列データの処理方法をとりあげる。利用するエンジンの処理として、例えば流通系企業のデータを導入すると、顧客情報、商品情報、POS 情報などが考えられるが、ここで、重要な機能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。これ以上の情報の導入は、時間軸に沿った解析によりその「時系列」の特有の意味を抽出する、という処理が必要である。そこで、時系列データに対する処理の一つとして、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。

2 生データの解釈

流通系のデータを列挙すると、顧客仕様の情報を、距離、購買日時を除く季節、天気、曜日、時帯というように、存在するデータは抽出しようとする知識の形態に応じて様々な解釈が可能である。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、あるいはデータとの関係を、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、より良質の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。

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

推論エンジンは、この入力から、TRANSACTION のクラスタリング ([3])、ITEM 間の関係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、時系列データを量子化する手段は、通常考えられる。実数の数値 MB に及ぶ流通業のデータを複数解析した結果、前述の 1 的例のように時間に 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」という観点でタイムスタンプを解析し、量子化する方法を検討する。

ここでは、以下的方法により、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 項の認識アルゴリズムにより、1 から ITEM, TIMESTAM P \rightarrow TIMING の参照表を得る。
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。

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る。時間帯の連続値の低音化とともに、そうした付加情報の変動度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、発生刻印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低音化を行なった。

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 情報を探すもののみのデータとの比較から、有用性は確認された。

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、ドメインに適した低音化単位の発見等が挙げられる。

参考文献

