データマイニングにおける最適区間相関ルールの効率的発見を支援するシステムの開発

A Data Mining System for Supporting Efficient Discovery of Association Rules with Optimized Ranges

胡　振江
Zhenjiang HU

東京大学大学院 情報理工学研究科

（〒113-8656 東京都文京区本郷 7-3-1　E-mail: hu@ipl.t.u-tokyo.ac.jp）

ABSTRACT. The general goal of data mining is to extract interesting correlated information from large collection of data. A key computationally intensive sub-problem of data mining involves mining association rules for market basket analysis. To help extracting useful association rules, we implemented a data mining system, which not only unifies the two algorithms (the standard Apriori algorithm and our new algorithm proposed recently) to deal with different size of database in a more efficient way, but also allows one to specify involved conditions to mine association rules with optimized ranges of his interest. We have test our system using the POS database from a coffee shop, and the experimental results indicate that out system is both fast and practical.

１．背景と目的

近年，バーコードやクレジットカードなどのデータ収集技術の大幅な進歩と，記憶装置の劇的な低価格化によって，非常に膨大なデータを蓄積することができるようになった．このようにして蓄積された膨大なデータから，例えば「目玉商品 A を購入した顧客は高い確信度で日用品 B を購入する」という規則が得られたとすれば，目玉商品 A が他のどの日用品 B の売上に貢献するかが明瞭になる．さらに目玉商品 A を購入した顧客の何パーセントが日用品 B を購入するかが分かれば，売上の予測にもある程度つながる．このように，膨大なデータの中から，無関係に見えそうなデータ間の関係を見つけだしたり有用な情報だけを選り分けたりする，データマイニングの理論や技術への需要がここ数年大きくなってきている．

「目玉商品 A を購入した顧客は高い確信度で日用品 B を購入する」という規則を

(目玉商品 A = yes) ⇒ (日用品 B = yes)

と表現する．一般に，A，B，C，D，…をあるデータベースの項目とするとき，

(A = a) AND (B = b) ⇒ (C = c) AND (D = d)

という形をした規則を相関ルール (association rule) と呼ぶ．また，相関ルールの左辺を「前項条件」，右辺を「後項条件」と呼ぶ．このような相関ルールの価値を判定する基準として，サポート (support)，コンフィデンス (confidence)，ゲイン (gain) の 3 つが存在する．

- サポート: 前項条件と後項条件とを同時に満たすデータの，全データに対する割合．
- コンフィデンス: 前項条件を満たすデータのうち，後項条件も同時に満たすデータの割合．
- ゲイン: 前項条件と後項条件とを同時に満たすデータの割合と，前項条件の割合に最小のコンフィデンスとして与えた数値をかけ合わせたものとの差．

このような相関ルールは，一般的に離散値をもつ属性を扱ったものである．しかしながら，現実のデータベースでは通常，年齢や預金残高のような数値属性が存在する．これらの数値属性を，離散値をもつ属性として考えても良いが，数値間の順序も考慮した結合ルールがあれば有用である．このような相関ルールに，最適区間相関ルールというものがある．最適区間相関ルールとは，数の属性を含むルールの中である条件を満たし，数の属性の区間が最長（すなわちサポートが最大)となるような相関ルールを抽出するものである．最適区間相関ルールの具体的な問題の例として，「パンを買った顧客がバターを買う相関ルールの中で，サポートが 50%から 70%の間，コンフィデンスが 30%以上，ゲインが 90%以上という条件のもと，最長の顧客年齢区間を求める」などがある．

従来の手法では発見される相関ルールは数が多く，その中からどのようにして有意義なものを見出すのかが問題とされている．さらに，既存の手法のみでは「サポート，コンフィデンス，ゲインについて，それらがある閾値以上である」という単純な条件のときしか効率的に最

適区間相関ルールを抽出することができないため，要求に充分に応えていないのが現状である．

このような問題を解決するため，本研究開発では，適切な条件を記述して対話的に最適区間相関ルールを高速に絞り込むためのシステムを開発した．本システムの具体的な目的は，大きく次の二つである．

(1) 相関ルールを抽出するアルゴリズムの研究

大規模なデータベースに対して効率よく動作する既存の Apriori アルゴリズムと，中規模のデータベースに対して高速に動作する我々のアルゴリズムとを実装し，本システムに統合することを目的とする．

(2) 最適区間相関ルールの研究

本研究開発ではより複雑な,「サポート, コンフィデンス，ゲインの任意の組み合わせと任意の範囲」によって記述された条件に対して，条件を満たす最適区間相関ルールを効率的に発見するための支援システムの開発を目的とする．

2．技術開発の内容

本研究開発では，まず，相関ルールを抽出する二つのアルゴリズム── (1) 大規模なデータベースに対して効率の良い既存の Apriori アルゴリズムと，(2) 中規模なデータベースに対して考案した Apriori アルゴリズムよりはるかに速いアルゴリズム──を実現し，1 つのシステムに統合する．次に，相関ルールの中から，サポート，コンフィデンス，ゲインによる任意の組み合わせ，任意の範囲で記述された条件で，効率的に最適区間相関ルールを抽出するためのアルゴリズムを実現する．最後に，これらのサービスを提供するサーバと，サービスを利用するクライアントからなるシステムを構築し，実際にネットワーク上で実行できる環境をつくる．

本研究開発のうち主要な技術課題は，今までに我々が提案した二つのアルゴリズム:

- 相関ルールの抽出アルゴリズム[1,2]
- 最適区間相関ルールの抽出アルゴリズム[3,4]

の実現とそれらの有効性を実証することである．

（1） 相関ルールの抽出アルゴリズム実現

我々は，プログラム運算手法に基づいて，相関ルールを抽出するための新しい効率的なアルゴリズム[1, 2]の導出に成功している．その成果は本研究開発の核となる理論である．これまでに我々が提案したアルゴリズムは，関数型言語 Haskell を用いて実現したが，実際に使えるシステムを作るには Java のような効率の良い言語で実装しなければならない．また，既存のアルゴリズムとの性能比較や，実際のデータ解析への適用はまだなされていない．

本研究開発の一つ目の技術課題は，相関ルールを抽出する二つのアルゴリズム──大規模なデータベースに対して効率的な既存の Apriori アルゴリズムと，中規模なデータベースに対して効率的な我々が提案した高速アルゴリズム──を統合して，データベースのサイズに基づいて抽出アルゴリズムを選択できるようなシステムを実現することである．

（2） 最適区間相関ルールの抽出アルゴリズムの実現

二つ目の技術課題は，サポート，コンフィデンス，ゲインに任意の組み合わせと，それらの任意の範囲で記述された条件に対して，効率的に最適区間相関ルールを抽出するアルゴリズムを実現することである．このアルゴリズムの実現のため「最小属性を持つ多次元探索木」の設計が重要である．また，使いやすいインターフェースの設計も重要である．

我々は，従来のものより一般的な最適区間相関ルールを抽出するための問い合わせ言語の定義と，効率的な実現手法を提案し[3]，その成果に評価を得ている．この成果は本研究開発の核となる理論である．また，本アルゴリズムの実現に重要な「最小属性を持つ多次元探索木」の理論については文献[4]を基礎としている．

3．システムの開発・実装

本システムは，サーバ，クライアント，および，サーバとクライアントの通信という三つの部分から構成される．サーバ側では，一般的な相関ルールを効率的に抽出するアルゴリズムを実現する機能，最適区間相関ルールを抽出する機能，および，テストデータの生成とシステム評価を行う機能から構成される．クライアント側では，サーバに実装された機能を利用するための相関ルール抽出のためのインターフェース，最適区間相関ルール抽出のためのインターフェース，および，解析結果の表示機能より構成される．サーバとクライアントの通信は，サーバとクライアント間のプロトコルおよびデータの交換機能から構成される．

具体的には，本システムは相関ルールを抽出するソフトウェア開発を支援するシステムである ARMiner[5]に，Java 言語を用いて機能を拡張することで実現した.各種のデータベースに対応できるよう，ARMiner 独自形式のファイルだけでなく，CSV (Comma-Separated-Value) 形式のファイルからもデータを読み込めるようになっている．これらのデータの流れに関する全体図を図1に示す．

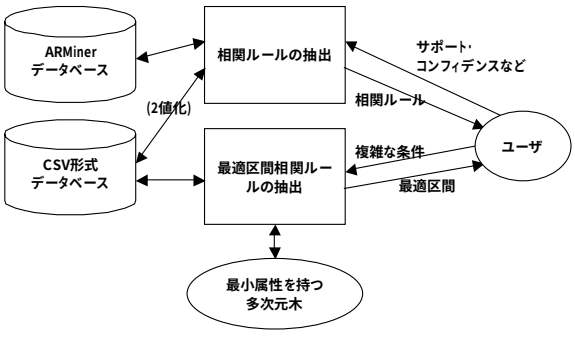


図 1 本システムのデータの流れに関する全体図

本システムは次の 4 つの主要機能から構成される．これらの機能に関する全体図を図 2 に示す．

(1) 相関ルールの効率的な抽出機能（S1，C1）
(2) 最適区間相関ルールの抽出機能（S2，C2）
(3) システム性能評価機能（S3，C3）
(4) サーバとクライアントとの連携機能（SC）

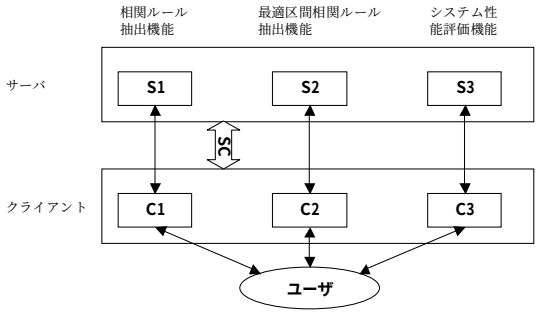


図 2 本システムの機能に関する全体図

（1）相関ルールの効率的な抽出機能

本機能は，サーバ上の機能である「相関ルール抽出機能」と，クライアント上の機能である「相関ルールの抽出のためのユーザインターフェース機能」から構成される．「相関ルール抽出機能」は，データベースをレコード毎に処理する方式と，処理効率の良いデータベースの属性を考慮した処理方式の両方をサポートする．なお，途中結果の確認のため，頻出集合も出力する．本機能を用いて，CSV 形式の顧客情報データベースから相関ルールを抽出する様子を図 3～図 5 に示す．

図 3 相関ルール抽出の条件設定画面

図 4 CSV 形式のデータを二値化する画面

図 5 抽出された相関ルール

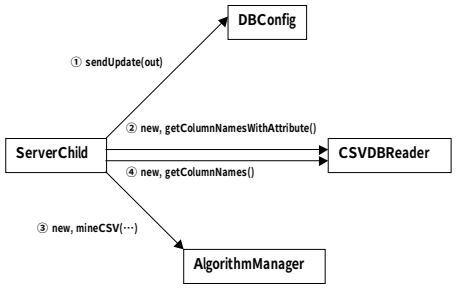
本機能は次のクラスから構成され，クラス間の関係は図 6.、図 7 のようになっている．

Aposteriori
論文[1，2]における相関ルール抽出の効率的なアルゴリズムを実装したクラス．ARMiner の仕様に従い，LargeItemsetsFinder インターフェースを実装している．

AlgorithmManager
ARMiner のもので，入力された条件下で相関ルールを抽出する機能の中心となるクラス．クライアントから指定されたサーバに登録済みのアルゴリズムの動的なロードもこのクラスのもとで行われる．

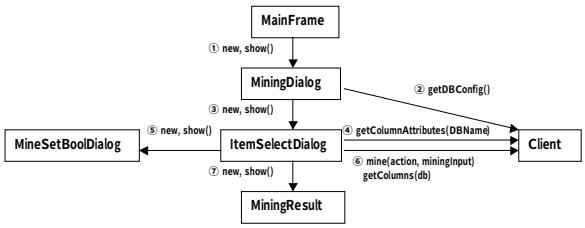
DatabaseReader
相関ルールを求める元となるデータベースを読み込むためのインターフェース．

CSVDBReader
CSV 形式のデータベースを読み込むクラス．CSV の形式については 3.3.節に詳細を述べる．



①：Client からの GETDBCONFIG メッセージによって呼び出し、DBConfig のコピーをClient に返す
②：Client からの GETCOLSWITHATTR メッセージによって呼び出す
③：Client からの MINECVS メッセージによって呼び出し、相関ルールを抽出して Client に返す
④：Client からの GETCOLS メッセージによって呼び出す

図 6「相関ルールの効果的な抽出」におけるクラス関係図



①：メニューから Find Association Rules を選ぶと呼び出し
②：new されたとき、利用できるアルゴリズム・データベース一覧を取得
③：Mine ボタンが押されると呼び出し
④：new されたとき、各項目の名前・属性を取得
⑤：数値属性の項目が選択される度に呼び出し
⑥：Execute ボタンが押されると呼び出し、実行結果を受け取る
⑦：実行結果が正しく受け取れたら呼び出し

図 7「相関ルールの効果的な抽出のためのユーザインターフェース」におけるクラス関係図

（2） 適区間相関ルールの抽出機能

本機能は，サーバ上の機能である「最適区間相関ルールの抽出機能」と，クライアント上の機能である「最適区間相関ルールの抽出のためのユーザインターフェース」から構成される．「最適区間相関ルールの抽出機能」で用いるアルゴリズムは，従来の単一的な条件に比べてより複雑な条件に対応できる．本機能を用いて，最適区間相関ルールを抽出する様子を図 8～図 9 に示す．

図 8 抽出された相関ルール

図 9 抽出された相関ルール

本機能は次のクラスから構成され，クラス間の関係は図 10 のようになっている．

CSVDBReader
CSV 形式のデータベースを読み込むクラス．

OptimizedRangesMiner
最適区間相関ルール抽出アルゴリズムが実装すべきインターフェース．

SimpleOptimizedRangesMiner
論文[3]による最適区間相関ルール抽出の効率的なアルゴリズムの実装のメインクラス．

Bucket
抽出する区間の最小単位幅でレコードを分類したデータ構造をあらわすクラス．

KdmTree
最小属性値を持つ多次元探索木を表現したクラス．論文[4]の実装である．

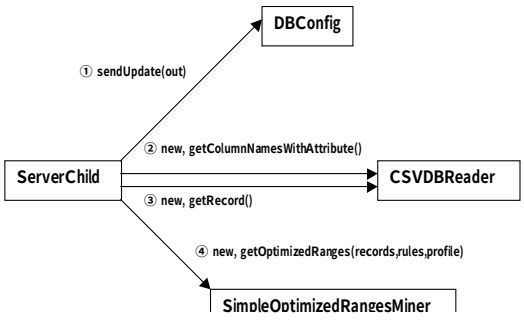


図 10「最適相関ルールの効果的な抽出」におけるクラス関係図

４．評価・考察

本システムに実装した評価機能を利用して，我々の提案した相関ルールの抽出アルゴリズム[1, 2] の性能評価を行った．

本システムの評価には ARMiner に付属しているデータベースを用い，サポートの値を 0.1～0.9 と変化させて実験した．実験の結果を図 13 に示す．

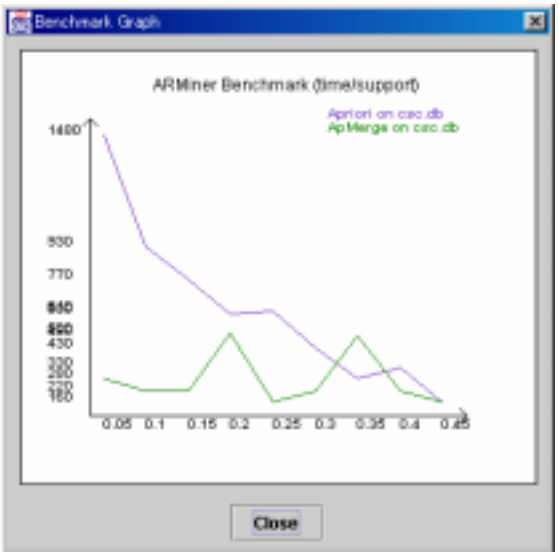


図 11 相関ルールの抽出機能の性能評価

この結果より，本システムで実装したアルゴリズム (Aposteriori) の方が標準的な Apriori アルゴリズムよりも高速に計算可能であることが確認された．

また，本システムに実装した性能評価機能に基づいて，我々の最適区間相関ルールの抽出アルゴリズム[3]の性能評価を行った．この評価には住宅情報のデータベース (700 レコード) を用い，項目「家賃」に対して Bucket のサイズを 100 刻みに，Tuple のサイズを 1～5 にして行った．評価実験の結果を図 14 に示す．

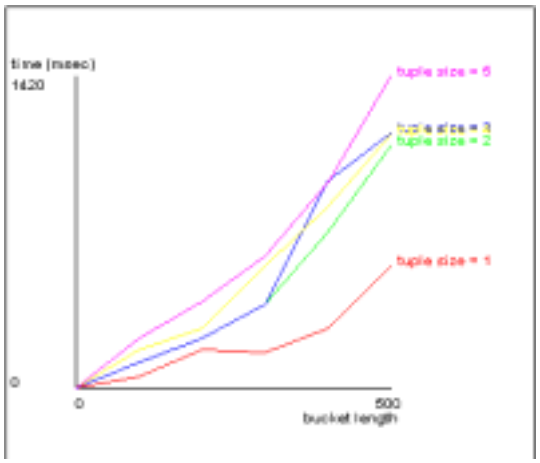


図 12 最適区間相関ルールの抽出機能の性能評価

この結果より，Bucket または Tuple を大きくすると計算時間が大きくなるが，実用的な範囲では，十分に効率的に動作していると判断することができる．

５．終わりに

本研究開発では，(1) 既存の相関ルールアルゴリズムと中規模なデータベースに対して高速な我々のアルゴリズムの統合，(2) 既存のアルゴリズムよりも複雑な条件 (サポート，コンフィデンス，ゲインによる任意の組み合わせ，任意の範囲によって記述される) で効率よく最適区間相関ルールを抽出できるアルゴリズムの実現を行った．また，これらのサービスを提供するサーバと，サービスと利用するクライアントからなるシステムを構築し，実際にネットワーク上で利用できる環境を実装した．

相関ルールの抽出については，既存の Apriori アルゴリズムと性能評価実験を行い，我々が提案したアルゴリズムの方が高速に相関ルールを抽出可能であることが確認できた．また，最適区間相関ルールのアルゴリズムについては，従来のアルゴリズムよりも複雑な条件のもとでも，実用的な時間で最適区間相関ルールを抽出できることが確認できた．これにより，膨大に蓄積され続けている情報を有効に活用し，より有意義な科学的発見を支援することが期待できる．

６．謝辞

本研究は，情報処理振興事業協会 (IPA) の委託により財団法人ソフトウェア工学研究財団 (RISE) が実施した平成 13 年度「高度情報化支援ソフトウェアシーズ育成事業」に採択され，株式会社東大総研と共同で開発したものです．ご支援に対して深く感謝いたします．

７．参考文献

[1] Z. Hu, W. N. Chin, M. Takeichi: “Calculating a New Data Mining Algorithm for Market Basket Analysis,” *Second International Workshop on Practical Aspects of Declarative Languages (PADL '00)*, Boston, Massachusetts, January 17-18, 2000.

[2] Z. Hu, W. N. Chin, M. Takeichi: “Calculating a New Data Mining Algorithm for Market Basket Analysis”, to appear in *Journal of Functional and Logic Programming*, MIT Press.

[3] H. Zhao, Y. Yokoyama, Z. Hu, M. Takeichi: “Mining Optimized Ranges: A Functional Approach,” *1st*

International Workshop on Programming and Programming Languages (IWPPL 2000), Singapore, December 17-19, 2000.

[4] H. Zhao, Z. Hu, M. Takeichi: "Multidimensional Searching Trees with Minimum Attribute," *Journal of JSSST Computer Software.*

[5] ARMiner Project, http://www.cs.umb.edu/~laur/ARMiner/