2C2-04

統計的相関性を考慮した局所的な特徴抽出の提案

Extract Local Correlation with Statistical Measure

谷口剛 Tsuyoshi TANIGUCHI
原口誠 Makoto HARAGUCHI

北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology Hokkaido University

In this paper, we propose an approach extracting local correlation between items from large databases. Local correlation is the only feature under some condition in databases. We represent a feature in databases using association rule. But the relation representation between items using association rule has some programs. The program of association rule is not considered correlation between items. Considering local correlation is solving two programs in association rule mining. First, correlation between items can be considered using statistical measure. Second, extracting local correlation utilize negative correlation that is not used in traditional correlation work. We believe this approach extracting local correlation contributes datamining research area.

1. はじめに

近年注目されている研究領域として、データマイニングの研究を挙げる事ができる。データマイニングの定義としては色々な定義があるが、大規模なデータから思いがけないパターンを発見すること [6]、が代表的な定義である。近年、マーケット領域や Web、医療領域など大規模なデータを抱える領域が増えてきた事などから、この実用技術が扱われる機会が非常に多くなってきている。

本研究では大規模なデータベースからの局所的な特徴抽出について議論していく。局所的な特徴とは、与えられたデータベースにおいてある条件を与えた時に、その条件で特に成立するような特徴を表す。本研究ではこの局所的な特徴を抽出する事により、データマイニングの研究領域において貢献できると考えている。それでは局所的な特徴を抽出する事にどのような意味があるか、次のような例題を考えていこう。ある全国展開しているチェーン店における全国各地にあるそれぞれの自店のデータベースを 1 つに統合したデータベースを考える。このデータから知識を得たい場合、単純にデータベース全体から得られる知識だけでなく、ある地域や年齢層などデータベースの部分に注目する事により、細やかな解析を行いたい場合もあるだろう。例えば、北海道地方の傾向が知りたい場合は、北海道を条件として解析を行う。しかしそのような解析では、北海道特有の傾向をとらえる事はできない。つまり、その解析でとらえた特徴は全国で成立しているような特徴も多く含むのである。例えば鳥インフルエンザの問題で全国的に鶏肉の売り上げが下がっていたとする。この傾向を北海道地方の解析によって発見しても、それは全国的な特徴であり、わざわざ北海道地方で限定してまで見つける必要のない情報であると本研究では考えている。したがってこの例題で言えば北海道で特に成立するような特徴を局所的な特徴としてとらえる。本研究では、条件付けによって特徴を抽出するのであれば、このような局所的な特徴をとらえる事が重要であると考えている。

さらに局所的な特徴をとらえる事により、従来手法では特徴抽出の妨げになる事が多かった高頻度のアイテムや、統計的に負の相関を持ったアイテム間の関係から、興味深い可能性の高い関係を抽出する事ができる。

連絡先: 谷口剛, 北海道大学大学院情報科学研究科, 〒 060-0814 札幌市北区北 14 条西 9 丁目, TEL(FAX):011-706-7161, E-mail:tsuyoshi@db-ei.eng.hokudai.ac.jp

2. 準備

この節では局所的な特徴を議論するための各種の定義を行う。

本研究では特徴を表現する方法として、**相関ルール (*Association Rule*)**[1] を使用した。その理由は、以下に示す 2 つの理由からである。

1. 簡潔でわかりやすい表現形式
   $X \Rightarrow Y$ という素朴な形式でアイテム間の関係を表現するため、直感的にわかりやすく、仮説が立てやすい。

2. 広い応用領域
   トランザクションデータベースの高い応用性により、様々な問題への対応が期待できる。

それでは、相関ルールについて定義していく。

$I = \{i_1, i_2, \ldots, i_m\}$ を**アイテム (*item*)** という。**トランザクション (*transaction*)**T は $T \subseteq I$ のようなアイテム集合とし、それぞれ TID という識別子を持っている。**トランザクションデータベース**D はトランザクションの集合とする。アイテムの集合 $X(X \subseteq I)$ は**アイテム集合 (*itemset*)** という。もし $X \subseteq T$ ならば、T は X を**含む**という。

相関ルールは $X \Rightarrow Y$ という形で表現される。ここで $X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$ とする。相関ルールの評価は主に**支持度 (*support*)** と**確信度 (*confidence*)** で行う。相関ルールの支持度は、トランザクションデータベースのトランザクションが X も Y も含む割合であり、ルールの適用範囲を表す。確信度は、トランザクションデータベースの X を含むトランザクションが Y を含む割合であり、ルールの信頼度を表す。**相関ルール発見問題**ではユーザが決めた支持度と確信度の閾値である最小支持度と最小確信度を満たす全ての相関ルールを価値のある相関ルールと考える。

本研究では、アイテム間の関係を評価をする際に**集合強度 (*Collective Strength*)**[4, 5] を用いる事にした。その理由は次節で詳しく述べるが、まずは以下に簡単に示す。

1. 確信度では表現できない統計的相関性 (correlation) を表現し、さらにアイテム間の強い正の相関も表現できる。

2. データベーススキャンが最大で 2 回ですむような高い計算効率を誇るアルゴリズムが開発されている。

それでは、集合強度の定義を行う。
アイテム集合 I の集合強度 $C(I)$ は以下の式で定義される。

$$C(I) = \frac{1 - v(I)}{1 - E[v(I)]} \cdot \frac{E[v(I)]}{v(I)}$$

ここで、$v(I)$ は全てのトランザクションにおけるアイテム集合 I の *violation* の割合、*violation* は I の 1 つ以上のアイテムは含まれるが、1 つ以上のアイテムは含まれないトランザクション、$E[v(I)]$ は統計的独立を仮定した時の $v(I)$ の値を表す。
集合強度の値はそれぞれ、0 はアイテム集合 I が完全な負の相関の状態、1 は統計的独立の状態、∞ は完全な正の相関の状態にある事を表す。

3. 相関ルールの問題点

この節では、なぜアイテム間の関係を集合強度で表現する必要があるのかについて議論していく。
集合強度という尺度が提案された背景には、Brin 等によって指摘された相関ルールの問題点がある [3]。次のような例題を考えていこう。アイテム A と B を考える。A の支持度は 40%、B の支持度は 90%、そして、アイテム集合 AB の支持度は 30%とする。この場合の A と B の関係を相関ルールで評価すると

$$A \Rightarrow B(\text{支持度} : 30\%, \text{確信度} : 75\%)$$

となり、このルールの支持度、確信度の値がユーザが指定した閾値を満たしていれば、価値があるルールとして出力する。しかし、この例題において B の支持度は 90%であり、A と B が関係がない (統計的独立の) 場合にも、A に対する B の確信度の値は 90%は期待されるため、A と B の間には負の相関がある。つまり確信度は単なる条件付確率であり、統計的な相関性 (correlation) は考慮していないのである。Brin 等はこの問題点を解決するため、統計的に正確な尺度であると評価が高い χ^2 検定によって相関していると評価されたアイテムの組を**統計的相関ルール (*correlation rule*)**[3] として出力する手法を提案した。
Brin 等による問題提起を受けて、Aggarwal 等は集合強度という尺度を用いてアイテム間の相関性を評価する手法を提案した [4, 5]。Aggarwal 等が解決しようとした問題点は主に以下の 2 つの点である。

1. 主に χ^2 検定を用いた手法における計算量の膨大さ
2. 正の相関の正確な評価

ここでは、主に 2 番目の点について議論する。
アイテム間の相関性を表現する素朴な方法として、関連研究において引用される事の多い尺度として、**興味深さの尺度 (*Interest Measure*)**[2] が挙げられる。興味深さの尺度は、以下の式で表現される。

$$\text{興味深さの尺度} = \frac{\text{実際の値}}{\text{統計的独立を仮定したときの値}}$$

この尺度で、先の例題の A と B の関係を評価すると、

$$\frac{0.3}{0.4 \times 0.9} = 0.83 < 1$$

より、統計的独立を仮定したときの値よりも実際の値が小さいため、この関係を負の相関と評価し、統計的な観点から重要ではないと評価できる。しかし、この興味深さの尺度には問題点がある。次のような例題を考えてみよう。アイテム X と Y を考える。それぞれのアイテムの支持度は 90%であり、X と Y はデータベースにおいて完全に同じトランザクションに含まれているとする。この関係を興味深さの尺度で評価すると、

$$\frac{0.9}{0.9 \times 0.9} = 1.11$$

となり、完全に正の相関の状態にあるにもかかわらず、統計的独立の時の値である 1 よりも少しだけ大きい値にしかならない。この関係を集合強度で評価すると、$v(I)$ の値が 0 になるため、集合強度の値は ∞ となり、興味深さの尺度と比べて正の相関を強く表現できる事がわかる。
本研究においてもアイテム間の関係は正確に表現したいと考えており、さらに局所的な特徴をとらえる際に、ある条件における正の相関を強く表現したいため、この尺度を用いてアイテム間の関係を評価した。

4. 局所的な特徴抽出

4.1 条件付相関

この節より、局所的な特徴について議論していく。まずは、局所的な特徴を考える前の準備として、相関と条件付相関について定義する。

定義 1 相関
アイテム集合 I を考える。$C(I) > \epsilon(\epsilon \geq 1)$ である時、I は**相関している**という。ここで、ϵ はユーザが決めた閾値とする。

つまり本研究でとらえる特徴とは相関しているアイテム集合である。次に、局所的相関を考える準備のために相関を利用して条件付相関を定義する。そのための尺度として条件付集合強度を定義する。

定義 2 条件付集合強度
アイテム集合 C、I を考える。C を含むトランザクションにおける I の集合強度を以下の式で定義する。

$$C(I|C) = \frac{1 - v(I|C)}{1 - E[v(I|C)]} \frac{E[v(I|C)]}{v(I|C)}$$

ここで、$v(I|C)$ は C を含むトランザクションにおける I の *violation* の割合、$E[v(I|C)]$ は統計的独立を仮定した時の $v(I|C)$ の値を表す。

条件付集合強度である条件についてのアイテム集合の評価を行う事ができる。条件付集合強度を用いて条件付相関を定義する。

定義 3 条件付相関
アイテム集合 C、I を考える。$C(I|C) > \epsilon(\epsilon \geq 1)$ である時、アイテム集合 I は**条件 C で相関している**という。ここで、ϵ はユーザが決めた値とする。

それでは、条件付相関を考える事によって何が嬉しいのか説明していこう。本研究では、条件付相関を考える事による利点を主に以下の 2 点であると考えている。

1. 相関性を考慮する従来手法では捨ててしまう関係の活用 (主に負の相関)

TID	itemset		
000	vegetable		
001	vegetable		
002	vegetable	meat	oil
003	vegetable	meat	oil
004	vegetable	dressing	
005	vegetable	dressing	
006	vegetable	dressing	
007	vegetable	oil	
008	oil		
009	meat		

図 1: 条件付相関の利点の例

2. 高頻度アイテムの活用 (主に多目的で利用されるアイテム)

ここでは主に 2 番目の利点について説明していこう。図 1 に示す例題で利点について説明していく。

この例題において、vegetable は野菜炒めや野菜サラダなど高頻度でさらに多目的で用いられるアイテムであると考える。実際のデータベースを考えても、1 つの目的のみで用いられるアイテムは稀であり、色々な用途で用いられるアイテムが多く存在する事が予想される。このようなアイテムを相関ルールで評価すると、結論部にそのアイテムが含まれるルールが多数出現し、ルールの可読性を低下させる原因となる。さらにルールの相関性を正確に表現すると、他のアイテムとの組み合わせの期待値が高くなるため、他のアイテムと負の相関の関係になりやすい。この例題においても、vegetable と meat を集合強度で評価すると、$C(\{vegetable, meat\}) = 0.70$ となり、重要ではない関係と評価される。しかし、条件付集合強度の閾値を 1 に設定してこのデータベースから条件付相関を導出すると、$C(\{vegetabel, meat\}|\{oil\}) = 3.00 > 1$ より、捨てるはずだった vegetable と meat の関係から、oil という条件における vegetable と meat の関係、つまり oil を条件に考えると野菜炒めをするために vegetable と meat の相関が上がるのを見つける事ができる。

4.2 局所付相関

前節で条件付相関の利点について議論したが、条件付相関には以下のような問題点がある。

- 全体のデータベースにおいて相関している関係が条件付相関にも含まれる事が多い

この事ははじめの例題でも議論したが、わざわざ条件づけて探さなくてもすでに全体の特徴としてとらえられている関係が、条件付相関に含まれてしまうのである。この問題点により、条件付相関の可読性が著しく低下してしまう事が予想される。したがって、局所的な相関を定義する事が必要になる。

定義 4 局所的相関
アイテム集合 C、I を考える。$C(I) < \epsilon(\epsilon \leq 1)$ かつ $C(I|C) > \delta(\delta \geq 1)$ である時、アイテム集合 I は**条件 C で局所的に相関している**という。ここで、ϵ、δ はユーザが決めた値とする。

局所的相関でとらえたい関係は要するに、全体のデータベースでは負の相関を持っているために重要な関係とはみなされないが、条件づける事により正の相関を持つような関係である。このように局所的相関を定義する事により、その条件で特に成立するような特徴も表現できるし、相関の変化もとらえる事ができる。さらに、条件付相関の冗長性も除去する事ができる。

4.3 アルゴリズム

本研究では局所的相関を導出するために、集合強度の性質を利用したアルゴリズムを使用した。集合強度の性質を以下に示す。

集合強度の性質
k_0 を 1 以上の数とし、アイテム集合 B のサイズを 2 以上とする。B の全ての 2-部分集合が k_0 以上の集合強度の値を持つならば、B は k_0 以上の集合強度の値を持つ

この性質の証明はほぼ実験的な証明だが、[4, 5] で与えられている。この性質を用いると、以下に示すアルゴリズムで相関しているアイテム集合を求める事ができる。

1. 与えられたデータベースから集合強度の閾値を満たす全ての 2-アイテム集合を導出する。
2. 1 で導出したアイテム集合を組み合わせて、全ての 2-部分集合が集合強度の閾値を満たす上位のアイテム集合を導出する。

つまり、2-部分集合が集合強度の閾値を満たすならば、そのアイテム集合の集合強度は閾値を満たすという性質を利用し、集合強度の閾値を満たす 2-アイテム集合を導出し、2-部分集合が全て集合強度の閾値を満たすような組み合わせを導出すれば、そのアイテム集合は集合強度の閾値を満たすのである。したがって、サイズが 3 以上のアイテム集合においてはデータベースをスキャンする必要はなく、1 回のスキャンを行うだけで相関しているアイテム集合を導出できる。

上記のアルゴリズムを利用し、相関しているアイテム集合を導出する事はできるが、本研究で導出したいのは局所的に相関しているアイテム集合である。与えられたデータベースにおいて局所的に相関しているアイテム集合を求める方法としては、大きく以下の 2 つの方法が考えられる。

1. 全体のデータベースで相関していないアイテム集合を生成し、そのアイテム集合の中で求めたい条件において相関しているアイテム集合を抽出する。
2. 全体のデータベースで相関しているアイテム集合と、求めたい条件において相関しているアイテム集合を生成し、求めたい条件において相関しているアイテム集合の中の全体のデータベースで相関しているアイテム集合を除いたアイテム集合を抽出する。

上記の方法の比較については、実験において議論する。それぞれのアイテム集合を求める段階においては、集合強度のアルゴリズムを利用している。

5. 実験

本研究の有効性とルール導出の実行時間を調べるため、実装実験を行った。用いた実験データは、UCI KDD Archive (http://kdd.ics.uci.edu) に集められているデータベースの中から、Entree Chicago Recommendation Data を使用した。このデータは、Entree Chicago レストラン推薦システムによって 1996 年 4 月から 1999 年 3 月にかけて集められた記録である。

地区	条件付相関	局所的相関	実行時間 (s)
Atlanta	5210	843	11.672
Boston	57757	3579	364.313
Chicago	26809	1171	265.625
Los Angeles	19699	1457	108.875
New Orleans	21186	1428	166.781
New York	11283	621	52.641
San Francisco	17314	779	100.391
Washington DC	99272	2300	917.969
計	258530	12178	1988.267

図 2: 相関しているアイテム集合を用いて局所的相関を求めた実験

このデータのデータ形式はトランザクションデータである。このデータの中から、Atlanta、Boston、Chicago、Los Angeles、New Orleans、New York、San Francisco、Washington DC という 8 つの都市のレストランの特徴を示すデータを使用した。したがって本実験において、これらの 8 つの地区を条件に局所的相関を抽出する事になる。実験環境は、CPU Pentium4 1.5GHz、RAM 384MB のスペックを持つ PC 上で行った。実装プログラムは、アプリケーション Microsoft Visual C++6.0 を使用し、C 言語で記述した。

まずは、本研究の有効性を調べるために局所的相関の内容分析を行った。抽出された関係の中にはいくつか興味深いものがあったが、Chicago 地方で見つかった以下の関係について考える。

Place for Singles　　Singles Scene
支持度:2.22%　　集合強度:2.44

この関係は局所的相関の性質より、全体のデータベースでは重要な関係ではないが、Chicago 地方で相関が高くなる関係である。したがってこの関係より、Chicago 地方の特徴として独身をターゲットにした店がいくつかあると予想する事ができる。

次に局所的相関を求める際の実行時間を調べる実験を行った。

まず図 2 に全体のデータベースにおいて相関しているアイテム集合と条件付相関のアイテム集合の比較により局所的相関を求める実験の結果を示す。この実験における条件は $\epsilon = 1.1$、用いた相関しているアイテム集合数は 26336 個、それを求めるための実行時間は 325.390s である。条件を $\epsilon = 1.1$ にした理由は、$\epsilon = 1.0$ に設定すると、計算量が膨大すぎて相関しているアイテム集合を導出する事が困難であったためである。これらの実験からわかる事は、今回のデータにおいては、条件付相関の 9 割以上がデータ全体でも相関しているアイテム集合であり、冗長である事が示された。さらに図 2 より、それぞれの都市の条件付相関の計算負担が実行時間に影響する事が示された。したがってこの方法では、条件が膨大になった時には対応できない。

次に、図 3 に全体で相関していないアイテム集合を利用して局所的相関を求める実験の結果を示す。この実験において、$\epsilon = 1.0$、用いた相関していないアイテム集合の数は 112218 個、それを求めるための実行時間は 718.172s である。この実験により、局所的相関導出のための実行時間は集合強度の値によらない事がわかった。この事は、探索が相関していないアイテム集合の範囲に絞られている事による。この方法によって全体のデータベースで相関しているアイテム集合と条件付相関の比較では出力する事ができなかった $\epsilon = 1.0$ における局所的相関を出力する事はできたが、全体の実行時間は早いとは言えず、効率の良いアルゴリズムの開発が課題として残った。

δ	1.2	1.1	1.0
Atlanta	21	76	580
Boston	9	57	700
Chicago	8	39	648
Los Angeles	16	56	570
New Orleans	24	74	488
New York	6	15	718
San Francisco	8	40	623
Washington DC	17	82	811
計	109	439	5138
時間 (s)	1030.89	1033.187	1030.844

図 3: 相関していないアイテム集合を用いて局所的相関を求めた実験

6. まとめ

本研究では、統計的な相関性を考慮して局所的な特徴を抽出するためのアプローチについて議論した。この手法により、ある条件だけで成立するような局所的な特徴をとらえる事ができる。今後の課題としては、別の尺度の選定、効率の良いアルゴリズムの開発なども挙げられるが、今回は決められた条件の下での局所的な特徴について議論していた。しかし、ユーザは局所的な特徴を与える条件を知りたい事も十分考えられ、本研究では局所的な特徴を与える条件の選定にも着手する予定である。

参考文献

[1] R.Agrawal , T.Imielinski , and A.Swami , "Mining Association Rules between Sets of Items in Large Databases" , *Proc. ACM SIGMOD Conf. Management of Data* , May , 1993 , p207-216.

[2] R.Srikant , and R.Agrawal , "Mining Quantitative Association Rules in Large Relational Tables" , *Proc. ACM SIGMOD Conf. Management of Data* , 1996.

[3] S.Brin , R.Motwani , and C.Silverstein , "Beyond Market Basket : Generalizing Association Rules to Correlations" , *Proc. ACM SIGMOD* , v26 , n2 , June , 1997 ,p265-276.

[4] C.C.Aggarwal , and P.S.Yu , "A New Framework for Itemset Generation" , *Proc. ACM Principles of Database Systems Conf.* , June , 1998 , p18-24.

[5] C.C.Aggarwal and P.S.Yu , "Mining Association with the Collective Strength Approach" , *IEEE Transactions on Knowledge and Data Engineering* , v13 , n6 , Nov-Dec , 2001 , p863-873.

[6] 福田剛志 , 森本康彦 , 徳山豪 , データマイニング , データサイエンスシリーズ , n3 , 共立出版 , 2001.